Dimensionality Reduction with Unsupervised Feature Selection and Applying Non-Euclidean Norms for Classification Accuracy

Amit Saxena, G G University, India
John Wang, Montclair State University, USA

ABSTRACT

This article presents a two-phase scheme to select reduced number of features from a dataset using Genetic Algorithm (GA) and testing the classification accuracy (CA) of the dataset with the reduced feature set. In the first phase of the proposed work, an unsupervised approach to select a subset of features is applied. GA is used to select stochastically reduced number of features with Sammon Error as the fitness function. Different subsets of features are obtained. In the second phase, each of the reduced features set is applied to test the CA of the dataset. The CA of a data set is validated using supervised k-nearest neighbor (k-nn) algorithm. The novelty of the proposed scheme is that each reduced feature set obtained in the first phase is investigated for CA using the k-nn classification with different Minkowski metric i.e. non-Euclidean norms instead of conventional Euclidean norm (L2). Final results are presented in the article with extensive simulations on seven real and one synthetic, data sets. It is revealed from the proposed investigation that taking different norms produces better CA and hence a scope for better feature subset selection.

Keywords: Classification, Data Mining, Feature Analysis, Genetic Algorithms, Minkowski Metric

INTRODUCTION

In most of the computer-based applications today, datasets are having a large number of patterns and relatively a smaller number of classes. Each pattern is characterized by a number of features and each pattern belongs to one of the total classes. Classification of these patterns is a major step in data mining (Han & Kamber, 2006). In majority of the data mining applications, patterns are required to be classified. As classification requires feature analysis, the latter becomes an important component of data mining. Feature analysis consists of
feature selection and feature extraction. The function of a feature selection process is to select a subset of features from the entire set of features in a dataset. Feature extraction on the other hand, may combine or re-compute features among themselves to create a new feature. Curse of dimensionality caused due to redundancy of extra or derogatory features is a major issue of concern in data mining. Feature selection process can be useful to counter curse of dimensionality problem. Feature selection is applied to select most significant features in a dataset. By significant features here, we mean those features, which alone can predict the classes of the patterns in a dataset with maximum accuracy. If the feature selection makes use of information (such as class of a pattern) given before the process is applied, then the approach is called supervised. If no information is supplied a priori to grouping the patterns, the approach is called an unsupervised. In later case, the features are combined on the basis of some similarity (such as clustering). A number of supervised feature selection methods exist which use Neural Networks, Fuzzy logic, k-nearest neighbor search (k-nn) algorithms. On the contrary, the problem of unsupervised feature selection has been addressed rarely.

In most of the unsupervised feature selection approaches, grouping of features is based on the distance among individual features with each other. The computation of distance is central in the unsupervised approaches to decide the level of similarity among features. The decision of selecting significant features is heuristic. To select most effective features from a large number of features, evolutionary computing techniques can be applied. Genetic algorithm (GA) (Romero & Abelló, 2009; Goldberg, 1989), is a powerful evolutionary computing technique based on the principles of evolution. GA can be applied to select features in this manner.

With reduced number of feature selected through GA, next essential objective is to test the classification accuracy (CA) of the dataset due to this subset of features. The k-nn classification is a supervised method used to determine the CA of a data set. The distance between the test pattern and each pattern in the dataset is determined and the class is decided on the basis of the class of the pattern having minimum distance from the test pattern. In most cases, k=1, i.e. the pattern having minimum distance from the test pattern is marked as the class of the test pattern. A popularly known distance used for this purpose is a Euclidean distance, which is a special case of Minkowskii metric or non-Euclidean norms. In this article, we vary the Minkowski metric parameters for different values including the popular Euclidean distance to observe the effect of variation of Minkowski metric on CA of the dataset due to a particular feature set. The article is organized as follows. Next section presents review of earlier work in the field. Section Genetic Algorithm, highlights brief description of GA. After that we outline Minkowski metric. The proposed method is explained in next. The brief summary of datasets is presented in Section Datasets. Simulation studies and Result Analysis are described separately in Appendix. Conclusions and future research scopes are presented in last section of main text.

REVIEW OF EARLIER WORK

Data mining is defined as a multidisciplinary joint effort from databases, machine learning, and statistics, which is championing in turning mountains of data into nuggets (Mitra, Murthy, & Pal, 2002). Feature selection (FS) is an important component of data mining as it helps in reducing dimensions of data set but still preserving the structure of the data set. Ho (1998) combined and constructed multiple classifiers using randomly selected features, which can achieve better performance in classification than using the complete set of features. As an option, the selection of an optimal feature vector can be made with an exhaustive search of all possible subsets of features (Zhang, 2005). FS methods can have two categories: the wrapper method; where the classification accuracy is employed to evaluate feature subsets and the
Related Content

An Innovation Ecosystem beyond the Triple Helix Model: The Trentino’s Case
Alberto Ferraris, Stefano Leucci, Stefano Bresciani and Fausto Giunchiglia (2016).
*Big Data: Concepts, Methodologies, Tools, and Applications* (pp. 1814-1831).
[www.igi-global.com/chapter/an-innovation-ecosystem-beyond-the-triple-helix-model/150244?camid=4v1a](www.igi-global.com/chapter/an-innovation-ecosystem-beyond-the-triple-helix-model/150244?camid=4v1a)

Fuzzy Miner: Extracting Fuzzy Rules from Numerical Patterns
*International Journal of Data Warehousing and Mining* (pp. 57-81).
[www.igi-global.com/article/fuzzy-miner-extracting-fuzzy-rules/1748?camid=4v1a](www.igi-global.com/article/fuzzy-miner-extracting-fuzzy-rules/1748?camid=4v1a)
Spatial OLAP and Map Generalization: Model and Algebra
Sandro Bimonte, Michela Bertolotto, Jérôme Gensel and Omar Boussaid (2012).
*International Journal of Data Warehousing and Mining* (pp. 24-51).
www.igi-global.com/article/spatial-olap-map-generalization/61423?camid=4v1a

Fusion Cubes: Towards Self-Service Business Intelligence
www.igi-global.com/article/fusion-cubes-towards-self-service/78287?camid=4v1a