SUPERVISED GROWING NEURAL GAS ALGORITHM IN CLUSTERING ANALYSIS

Apirak Jirayusakul

A Dissertation Submitted in Partial
Fulfillment of the Requirements for the Degree of
Doctor of Philosophy (Computer Science)
School of Applied Statistics
National Institute of Development Administration
2007

SUPERVISED GROWING NEURAL GAS ALGORITHM IN CLUSTERING ANALYSIS

Apirak Jirayusakul

School of Applied Statistics

The Examining Committee Approved This Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy (Computer Science).

Associate Professor ....Committee Chairperson
(Surapong Auwatanamongkol, Ph.D.)

Associate Professor ....Committee
(Boontee Kruatrachue, Ph.D.)

Associate Professor ....Committee
(Pipat Hiranvanichakorn, Ph.D.)

Assistant Professor ....Committee
(Ohm Sornil , Ph.D.)

Associate Professor .... Dean
(Surapong Auwatanamongkol, Ph.D.)
Date ...Jan 18, 2008...

ABSTRACT

Title of Dissertation	Supervised Growing Neural Gas Algorithm in Clustering Analysis
Author	Apirak Jirayusakul
Degree	Doctor of Philosophy (Computer Science)
Year	2007

Nowadays, supervised clustering algorithms become important tools in many application areas, such as bioinformatics (Dettling and Bühlmann, 2002), pattern classification (Zeidat, 2005) and information retrieval (Finley and Joachims, 2005). The supervised clustering algorithms do not only group objects based on similarities of objects but also class labels of objects, so these algorithms can avoid grouping similar objects of different class labels into the same cluster. In this thesis, two supervised clustering algorithms based on a prototype-based paradigm are introduced, namely Supervised Growing Neural Gas algorithm (SGNG) and Robust Supervised Growing Neural Gas algorithm (RSGNG).

The SGNG algorithm is designed to utilize the class labels of training objects to guide a clustering process by incorporating several techniques such as the Type Two Learning Vector Quantization (LVQ2) (Kohonen, 1990) technique, the modified adaptive learning rate technique and the cluster repulsion mechanism of the Robust Growing Neural Gas algorithm (RGNG) (Qin and Suganthan, 2004). Furthermore, a novel prototype insertion mechanism is introduced to discover an unidentified cluster region.

The RSGNG algorithm is a modified version of the SGNG algorithm that possesses robust properties such as the insensitivities to the initialization of prototypes, the order of training inputs and the presence of noise and outliers. The RSGNG algorithm combines the SGNG learning schema with an outlier resistant strategy. In addition, the prototype insertion mechanism of the SGNG algorithm is modified to eliminate the influences of noise and outliers.

The two novel validity indices for supervised clustering, namely $\mathit{Validity}(K)$ and $\mathit{MDL}(K)$, are introduced as the tools for determining an optimal clustering that yields low intra-distances, high inter-distances, and low cluster impurities. The $\mathit{Validity}(K)$ index determines an optimal clustering based on cluster geometry measurements (the intra-distances between the members of any cluster and the inter-distances between prototypes) and cluster impurity measurement. The $\mathit{MDL}(K)$ index, which is based on an information theory paradigm, modifies the unsupervised Minimum Distance Length index (Qin and Suganthan, 2004; Bischof, Leonardis and Selb, 1999) by including cluster impurity measurement as another factor to determine the optimal number of clusters.

To evaluate the effectiveness of the SGNG and RSGNG algorithms, four experiments are conducted. The first experiment uses two sets of synthetic datasets, one with introduced noise and outliers, and the other without these, to illustrate graphically abilities of the two proposed algorithms with respect to their growing ability, their ability to cluster adjacent regions of different classes, and their robust properties. The experiment is also intended to evaluate the effectiveness of the $\mathit{Validity}(K)$ and $\mathit{MDL}(K)$ validity indices. Based on the results of this experiment, it can be seen that the SGNG and RSGNG algorithms using the two proposed validity indices give almost the same clustering results on every dataset without introduced noise and outliers. For the datasets with introduced noise and outliers, the RSGNG algorithm outperforms the SGNG algorithm. In addition, the $\mathit{MDL}(K)$ can determine the optimal number of clusters correctly while the $\mathit{Validity}(K)$ cannot.

In the second experiment, the robust properties of the SGNG and RSGNG algorithms are evaluated using UCI datasets. The $\mathit{MDL}(K)$ is used in this experiment since it performs better in the first experiment. Based on the results of this experiment confirm that both the SGNG and the RSGNG perform quite the same on the datasets

without introduced noise and outliers but the RSGNG performs better than the SGNG on the datasets with introduced noise and outliers.

In the third experiment, the performances of the SGNG and RSGNG algorithms are compared with those of other three supervised clustering algorithms (Zeidat, 2005) using some UCI datasets. The experimental results show that the two proposed algorithms perform better than the other three supervised clustering algorithms in terms of cluster impurities and total running times.

In the fourth experiment, two synthetic datasets containing complex shape clusters are used to study the topological formulation ability for the original GNG, the SGNG and the RSGNG, respectively. The experimental results show that the SGNG and RSGNG algorithms, when compared with the original GNG, have comparable abilities of topological formulation and can even form cluster regions with less misclassification rates and fewer numbers of prototypes.

ACKNOWLEDGEMENTS

Many people have provided advice, support, and encouragement to the author, during the research which led to this dissertation I would like to express my heartfelt appreciation to:

My supervisor, Assoc. Prof. Dr. Surapong Auwatanamongkol, for his generous support and intellectual guidance throughout my years as a graduate student; his insightful advice, clear vision, many suggestions, and endless efforts to be available for many educational discussions, were invaluable.

My co-advisor, Assoc. Prof. Dr. Boontee Kruatrachue (Department of Computer Engineering, King Mongkut's Institute of Technology Ladkrabang), whose encouragement was invaluable and who provided many valuable comments on drafts of this dissertation.

Dr. Aunnupan (Sitisak) Rodtook (Department of Computer Science, Ramkhamhaeng University) who shared valuable ideas and did help along the research.

Dr. Zen Harper, Assoc. Prof Dr.Pachara Chaisuriya (Department of Mathematics, Mahidol University), and Prof. Dr.Vasile Palade (Computing Laboratory Oxford University England) with helps for publishing the proposed paper (*A supervised growing neural gas algorithm for cluster analysis*) to The International Journal of Hybrid Intelligent Systems (IJHIS) Volume 4, Number 2.

Assoc. Prof. Dr. P.N. Suganthan and Dr. A.K. Qin (Nanyang Technological University,Singapore) for their RGNG source code and datasets, Dr. Nidal Zeidat for his Ph.D. Dissertation copy (Supervised Clustering: Algorithms and Application) and source code. In addition, Prof. Dr. Bernd Frizke (father of GNG) and Dr. Jim Holmström (http://www.it.uu.se/katalog/jimh) for publishing their GNG source codes on internet.

Mr.Kongvut Lertvorratham , Plt.off. Thanongsak Aneksrup and Mrs. Jiratta Phuboon-Ob, who shared with the pains and happiness during the course of this work; her endless support, sacrifice, and understanding kept me going through it all.

Finally, my mother, Mrs. Lamaipan Jirayusakul, who first taught me the importance of education, my sister Miss Sorawan for help with English, and especially, Miss Nantana Sawekwapee for taking care. I dedicate this dissertation to my parent.

Mr. Apirak Jirayusakul
26 Dec 2007