UNIVERSITY OF JOENSUU
COMPUTER SCIENCE
DISSERTATIONS 9

OLLI VIRMAJOKI

PAIRWISE NEAREST NEIGHBOR METHOD REVISITED

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Science of the University of Joensuu, for public criticism in the Louhela Auditorium of the Science Park, Länsikatu 15, Joensuu, on December 3rd, 2004, at 12 noon.

UNIVERSITY OF JOENSUU
2004

Supervisor Professor Pasi Fränti
Department of Computer Science
University of Joensuu
Joensuu, Finland

Reviewers Professor Martti Juhola
Department of Computer Sciences
University of Tampere
Tampere, Finland

Professor Olli Nevalainen
Department of Computer Science
University of Turku
Turku, Finland

Opponent Professor Jukka Teuhola
Department of Computer Science
University of Turku
Turku, Finland

ISBN 952-458-571-5
ISSN 1238-6944
Computing Reviews (1998) Classification: G.1.6, H.3.3, I.4.2, I.4.6, I.5.1, I.5.3, I.5.4
General Terms: Algorithms
Yliopistopaino
Joensuu 2004

Pairwise Nearest Neighbor Method Revisited

Olli Virmajoki

Department of Computer Science
University of Joensuu
P.O.Box 111, FIN-80101 Joensuu FINLAND
ovirma@cs.joensuu.fi

University of Joensuu, Computer Science, Dissertations 9
Joensuu, 2004, 164 pages
ISBN 952-458-571-5, ISSN 1238-6944

Abstract

The *pairwise nearest neighbor* (*PNN*) method, also known as *Ward's* method belongs to the class of agglomerative clustering methods. The *PNN* method generates hierarchical clustering using a sequence of merge operations until the desired number of clusters is obtained. This method selects the cluster pair to be merged so that it increases the given objective function value least.

The main drawback of the *PNN* method is its slowness because the time complexity of the fastest known exact implementation of the *PNN* method is lower bounded by $\Omega(N^2)$, where N is the number of data objects. We consider several speed-up methods for the *PNN* method in the first publication. These methods maintain the precision of the method. Another method for speeding-up the *PNN* method is investigated in the second publication, where we utilize a k-neighborhood graph for reducing distance calculations and operations. A remarkable speed-up is achieved at the cost of slight increase in distortion.

The *PNN* method can also be adapted for multilevel thresholding, which can be seen as a 1-dimensional special case of the clustering problem. In the third publication**,** we show how this can be implemented efficiently using only $O(N \cdot \log N)$ time, in comparison to a straightforward approach that requires $O(N^2)$.

The merge philosophy is extended, by using the iterative shrinking method, in the fourth publication. In the merge phase of the *PNN* method, the two nearest clusters are always joined. Instead of this, we assign data objects to the neighboring clusters that they belong to. In this way, we get better clustering results; however, the results come at the cost of an increase in the running time. The proposed method is also used as a crossover method in a genetic algorithm, which produces the best clustering results in respect of the minimization of intra cluster variance.

The *PNN* algorithm can also be applied to generating optimal clustering. In the fifth publication**,** we use a branch-and-bound technique for finding the best possible

clustering by generating a sequence of merge operations. Instead of using the local optimization strategy in the merge phase, we consider every possible merge by constructing a search tree, in which each merge performs the branch. We are also able to reduce the search space under certain bounding conditions. In addition, we give two polynomial time variants that utilize the proposed branch-and-bound technique, which only construct the search tree to a limited depth.

Keywords: agglomerative clustering, codebook generation, clustering algorithms, pairwise nearest neighbor method, pattern recognition, unsupervised learning, vector quantization, Ward's method.

Acknowledgements

The work presented in this thesis was carried out at the Department of Computer Science, University of Joensuu, Finland during 2000-2004 while the author also acted as a fulltime lecturer with the Kajaani Polytechnic.

I would like to express my sincere gratitude to my thesis supervisor, Professor Pasi Fränti, for his guidance, encouragement and infinite support throughout the research. I also owe my thanks for the cooperative work done by Dr. Timo Kaukoranta and Ville Hautamäki. I am thankful to Professor Martti Juhola and Professor Olli Nevalainen, the reviewers of this thesis, for their helpful comments and recommendations.

I would like to express my sincere thanks to the Department of Computer Science, University of Joensuu, for the grants for the conference trips, and to the East Finland graduate school for Computer Science and Engineering (ECSE) for organizing excellent summer schools.

I am grateful to all the people of Kajaani Polytechnic, Kajaani, for providing me support and the opportunity to work in this fruitful environment.

I owe my special thanks to my wife Riitta, to my daughters Noora and Roosa, and to my son Joonas for their endless support, encouragement, and understanding.

Finally, I would like to express my gratitude to any, anonymous, person who has made a positive contribution to my life.

Joensuu, November 2004

Olli Virmajoki

Abbreviations and symbols

Abbreviations

AESA	approximating and eliminating search algorithm
arg	argument
BB	branch-and-bound
CA	competitive agglomeration
CL	complete linkage
ECSE	East Finland graduate school for Computer Science and Engineering
GA	genetic algorithms
GAIS	genetic algorithm with the iterative shrinking as crossover
GLA	generalized Lloyd algorithm
IS	iterative shrinking
ISODATA	iterative self-organizing data analysis technique
KD-tree	*k*-dimensional tree
*k*NN graph	*k*-nearest neighbor graph
LBG	Linde, Buzo, and Gray
LMQ	Lloyd-Max quantizer
log	logarithm
MDL	minimum description length
MPS	mean-distance-ordered partial search
MSE	mean square error
NN	nearest neighbor
PDS	partial distortion search
PNN	pairwise nearest neighbor
RGB	red, green, blue
RLS	randomized local search
SAGA	self-adaptive genetic algorithm
SL	single linkage
SOM	self-organizing map
TIE	triangular inequality

Symbols

c	code vector
C	codebook: a set of code vectors $C=\{c_1, c_2, \ldots, c_M\}$
d	distance function
g	number of GLA iterations
k	number of nearest neighbors
K	dimension of vector
M	number of clusters / size of codebook
M_0	size of a preliminary codebook

N	number of data vectors
nn	nearest neighbor pointer
p	mapping to the partition
P	partition: a set of mappings $P=\{p_1, p_2, \ldots, p_N\}$
s	cluster
S	clustering: a set of clusters $S=\{s_1, s_2, \ldots, s_M\}$
x	data vector
X	a set of data vectors $X=\{x_1, x_2, \ldots, x_N\}$
τ	number of incoming pointers
O	asymptotic order
Ω	lower bound

List of original publications

P1. O. Virmajoki, P. Fränti and T. Kaukoranta, Practical methods for speeding-up the pairwise nearest neighbor method, *Optical Engineering*, 40(11): 2495-2504, November 2001.

P2. P. Fränti, O. Virmajoki and V. Hautamäki, Effective agglomerative clustering by *k* nearest neighbor graph. *Submitted*. Preliminary version has been published in [FVH03, VF04].

P3. O. Virmajoki and P. Fränti, Fast PNN based algorithm for multilevel thresholding, *Journal of Electronic Imaging*, 12(4): 648-659, October 2003.

P4. P. Fränti and O. Virmajoki, Iterative shrinking method for the clustering problems. *Submitted*. Preliminary version has been published in [VFK02, FV03].

P5. P. Fränti and O. Virmajoki, Optimal clustering by merge-based branch-and-bound. *Submitted*. Preliminary version has been published in [FVK02, FV02].

Contents

1 Introduction 1

2 Clustering algorithms 4

2.1 Algorithms 4
2.2 Unknown number of clusters 6
2.3 Fast search methods 7
2.4 Multilevel thresholding 8

3 Agglomerative clustering 10

3.1 PNN method 10
3.2 Using a distance matrix 15
3.3 Fast exact PNN 16
3.4 Lazy PNN 19
3.5 Inexact variants 20
3.6 Genetic algorithm 23
3.7 Summary 25

4 Summary of the publications 26

5 Summary of the results 29

6 Conclusions 33

7 References 34

Publications