ADVANCES ON INFORMATION PROCESSING AND MANAGEMENT

CONFERENCE ORGANIZERS INSTITUTES

The International Conference on Integrated Information is supported by the following Institutes:

Emerald Group Publishing Limited
Technological educational Institute of Athens, Greece
University of Peloponnese, Greece
National And Kapodistrian University of Athens, Greece
Mednet Hellas, The Greek Medical Network
2nd AMICUS Workshop

To learn more about I-DAS, including the Book Series, please visit the webpage
http://www.i-das.org/

INTEGRATED INFORMATION

International Conference on Integrated Information

Kos, Greece September, 29 – October, 3 2011

EDITORS

Georgios A. Giannakopoulos
Technological Educational Institute of Athens, Greece

Damianos P. Sakas
University of Peloponnese, Greece

All papers have been peer-reviewed

IDAS Institute for the Dissemination of Arts and Science

Piraeus, Greece, 2011

Editors

Georgios A. Giannakopoulos

Technological Educational Institute of Athens
Faculty of Management and Economics
Department of Library Science and Information Systems
Address: Aghiou Spyridonos Street, 12210, Egaleo
E-mail: gian@teiath.gr

Damianos P. Sakas

University of Peloponnese
Faculty of Science and Technology
Department of Computer Science and Technology
Address: End of Karaiskaki St., 22100, Tripolis, Greece
E-mail: D.Sakas@uop.gr

ISSN:

Printed in the Greece, EU

CONTENTS

Preface: Proceedings of the International Conference on Integrated Information (IC-ININFO 2011)

GEORGIOS A. GIANNAKOPOULOS

Department of Library Science and Information Systems, Technological Educational Institute of Athens, Aghiou Spyridonos, Egaleo, 12210, Greece

DAMIANOS P. SAKAS

Department of Computer and Technology Science, University of Peloponnese, Praxitelous 89-91, Piraeus, 18532, Greece

Aims and Scope of the Conference

The International Conference on Integrated Information 2011 took place in Kos Island, Greece, between September, 29 and October, 3, 2011. IC-ININFO is an international interdisciplinary conference covering research and development in the field of information management and integration.

The conference aims at creating a forum for further discussion for an Integrated Information Field incorporating a series of issues and/or related organizations that manage information in their everyday operations. Therefore, the call for papers is addressed to scholars and/ or professionals of the fields of Library and Archives Science (including digital libraries and electronic archives), Museum and Gallery Studies, Information Science, Documentation, Information Management, Records Management, Knowledge Management, Data management and Copyright experts the latter with an emphasis on Electronic Publications. Furthermore, papers focusing on issues of Cultural Heritage Management and Conservation Management are also be welcomed along with papers regarding the Management of Nonprofit Organizations such as libraries, archives and museums.

One of the primary objectives of the IC-ININFO will be the investigation of information-based managerial change in organizations. Driven by the fast-paced advances in the Information field, this change is characterized in terms of its impact on organizations that manage information in their everyday operations.

Grouping emerging technologies in the Information field together in a close examination of practices, problems and trends, IC-ININFO and its emphases on integration and management will present the state of the art in the field. Addressed jointly to the academic and practitioner, it will provide a forum for a number of perspectives based on either theoretical analyses or empirical case studies that will foster dialogue and exchange of ideas.

Topics of general Interest

Library Science, Archives Science, Museum and Gallery Studies, Information Science, Documentation, Digital Libraries, Electronic Archives, Information Management, Records / Document Management, Knowledge Management, Data Management, Copyright, Electronic Publications, Cultural Heritage Management, Conservation Management, Management of Nonprofit Organizations, History of Information, History of Collections, Health Information

Symposia

The Conference offered a number of sessions under its patronage, providing a concise overview of the most current issues and hands-on experience in information-related fields.

- Symposium on Integrated information: Theory, Policies, Tools
- 4th Symposium on Business and Management and Dynamic Simulation Models supporting management strategies

- Session on Open Access Rrepositories: Self-archiving, Metadata, Content policies, Usage
- Session on Evidence-Based Information in Clinical Practice
- Session on Business Management and Communication Strategies supporting Decision Making Process in Tourism Sector
- Session on Electronic Publishing: A Developing Landscape
- Session on Information and Knowledge Management
- Session on Information Content Preservation as Outcome of Conservation of Cultural Heritage: Ethics, Methodology and Tools
- Session on Advances Information for Strategic Management
- Session on Information History: Perspectives, Methods and Current Topics
- Session on Divergence and Convergence: Information Work in Digital Cultural Memory Institutions
- Session on Contemporary issues in Management: Organisational Behaviour, Information Technology, Education & Hospital leadership.

The wide range of aspects that the sessions covered, highlighted future trends in the Information Science.

Paper Peer Review

More than 300 papers had been submitted for consideration in IC-ININFO 2011. From them, 91 were selected for presentation, after peer review in a double blind review process. The accepted papers were presented at IC-ININFO 2011.

Thanks

We would like to thank all members that participated in any way in the IC-ININFO 2011 Conference and especially:

- The famous publishing house Emerald for its communication sponsorship.
- The co-organizing Universities and Institutes for their support and development of a high-quality Conference scientific level and profile.
- The members of the Scientific Committee that honored the Conference with their presence and provided a significant contribution to the review of papers as well as for their indications for the improvement of the Conference.
- All members of the Organizing Committee for their help, support and spirit participation before, during and after the Conference.
- The Session Organizers for their willing to organize sessions of high importance and for their editorial work, contributing in the development of valued services to the Conference.
- PhDc Marina Terzi for her excellent editorial work, contributing in the production of the Conference proceedings.

CONFERENCE DETAILS

Chairs

Georgios A. Giannakopoulos, Technological Educational Institute of Athens, Greece
Damianos P. Sakas, University of Peloponnese, Greece

Co-Chairs

Daphne Kyriaki – Manesi, Technological Educational Institute of Athens, Greece
Dimitrios Vlachos, University of Peloponnese, Greece

Scientific Committee

Amanda Spink, Queensland University of Technology
Andreas Bagias, European Court
Andreas Rauber, Vienna University of Technology
Astrid van Wesenbeeck, SPARC Europe
Christine Urquhart, Aberystwyth University
Christos Schizas, University of Cyprus
Christos Skourlas, Technological Educational Institute of Athens
Claire Farago, University of Colorado at Boulder
Claus-Peter Klas, FernUniversität in Hagen
Costas Vassilakis, University of Peloponnese,
Dimitris Dervos, Technological Educational Institute of Thessaloniki
Eelco Ferwerda, OAPEN
Elena Garcia Barriocanal, University of Alcalá
Emmanouel Garoufallou, Technological Educational Institute of Thessaloniki
Filippos Tsimpoglou, University of Cyprus
Fillia Makedon, University of Texas at Arlington
George Korres, University of Newcastle
Georgios Evangelidis, University of Macedonia
Georgios Panagiaris, Technological Educational Institute of Athens
Johan Oomen, Netherlands Institute for Sound and Vision
José Aldana, University of Malaga
Konstantinos Masselos, University of Peloponnese
Luciana Duranti, The University of British Columbia
Markos N. Dendrinos, Technological Institute of Athens
Milena Dobreva, University of Strathclyde
Prodromos Tsiavos, London School of Economics and Political Science
Sándor Darányi, University of Borås
Sarantos Kapidakis, Ionian University
Sirje Virkus, Tallinn University
Spiros Zervos, Technological Educational Institute of Athens
Susan Myburgh, University of South Australia
Theodoros Pitsios, University of Athens, Faculty of Medicine

Organizing Committee

Alexandros Koulouris (Chair), Technological Educational Institute of Athens
Christos Christopoulos, SCEV Scientific Events Ltd
Marina Terzi, University of the Aegean, Greece
Evangelia Markaki, Aristotle University of Thessaloniki

Assimina Kaniari, Athens School of Fine Arts
Evangelia Lappa, General Hospital Attikis K.A.T.
Dimitris Kouis, Greek Ministry of Education, Lifelong Learning and Religious Affairs
Dionysis Kokkinos, National Technical University of Athens

KEYNOTE SPEAKER

Professor Amanda Spink

Professor Amanda Spink has published over 340 scholarly journal articles, refereed conference papers and book chapters, and 6 books. Many of her journal articles are published in the Journal of the American Society for Information Science and Technology, Information Processing and Management, and the Journal of Documentation. She is Editor of the Emerald journal Aslib Proceedings. Amanda's research has been published at many conferences including ASIST, IEEE ITCC, CAIS, Internet Computing, ACM SIGIR, and ISIC Conferences. Her recent books include Information Behavior: An Evolutionary Instinct and Web Search: Multidisciplinary Perspectives, both published by Springer. Amanda's research focuses on theoretical and empirical studies of information behavior, including the evolutionary and developmental foundations. The National Science Foundation, the American Library Association, Andrew R. Mellon Foundation, Amazon.com, Vivisimo. Com, Infospace.com, NEC, IBM, Excite.com, AlltheWeb.com, AltaVista.com, FAST, and Lockheed Martin have sponsored her research. In 2008 Professor Spink had the second highest H-index citation score in her field from 1998 to 2008 [Norris, M. (2008)]. Ranking Fellow Scholars and their H-Index: Preliminary Survey Results. Loughborough University, Dept of Information Science Report].

An Extensive Experimental Study on the Cluster-based Reference Set Reduction for Speeding-up the k-NN Classifier

Stefanos Ougiaroglou[†], Georgios Evangelidis[†] and Dimitris A. Dervos[‡]

†University of Macedonia, Department of Applied Informatics, 54006, Thessaloniki, Greece
stoug, gevan (at) uom.gr

‡Alexander TEI of Thessaloniki, Department of Informatics, 57400, Sindos, Greece dad (at) it.teithe.gr

Abstract: *The k-Nearest Neighbor (k-NN) classification algorithm is one of the most widely-used lazy classifiers because of its simplicity and ease of implementation. It is considered to be an effective classifier and has many applications. However, its major drawback is that when sequential search is used to find the neighbors, it involves high computational cost. Speeding-up k-NN search is still an active research field. Hwang and Cho have recently proposed an adaptive cluster-based method for fast Nearest Neighbor searching. The effectiveness of this method is based on the adjustment of three parameters. However, the authors evaluated their method by setting specific parameter values and using only one dataset. In this paper, an extensive experimental study of this method is presented. The results, which are based on five real life datasets, illustrate that if the parameters of the method are carefully defined, one can achieve even better classification performance.*

Keywords: *K-NN classification, Clustering, Data reduction, Scalability*

I. INTRODUCTION

The data mining algorithms that assign new data items into one of a given number of categories (or classes) are called classifiers (Han and Kamber, 2000). Classifiers can be evaluated by two major criteria: classification accuracy and computational cost. k-NN is an extensively used and effective lazy classifier (Dasarathy, 1991). It works by searching the training data in order to find the k nearest neighbors to the unclassified item x according to a distance metric. Then, x is assigned into the most common class among the classes of the k nearest neighbors. Ties are resolved either by choosing the class of the one nearest neighbor or randomly. This work adopts the first approach.

However, the k-NN classifier has the major disadvantage of high computational cost as a consequence of the computations needed to estimate all distances between a new, unclassified, item and the training data. Thus, as the size of the training set becomes larger, the computational cost increases linearly.

Many researchers have focused on the reduction of the k-NN computational cost and therefore several speedup methods have been proposed. These methods are mainly based on either indexing (Samet, 2005; Zezula et al, 2006) or data reduction techniques (Wilson and Martinez, 2000; Lozano, 2007).Additionally to these methods, recent research proposed cluster-based approaches for speeding-up the k-NN classifier, such as, the Cluster-based Trees (Zhang and Srihari, 2004), the Representative-based Supervised Clustering Algorithms (Eick et al, 2004), and, the Reference Set Reduction method through k-means clustering (Hwang and Cho, 2007). This work focuses on the latter approach.

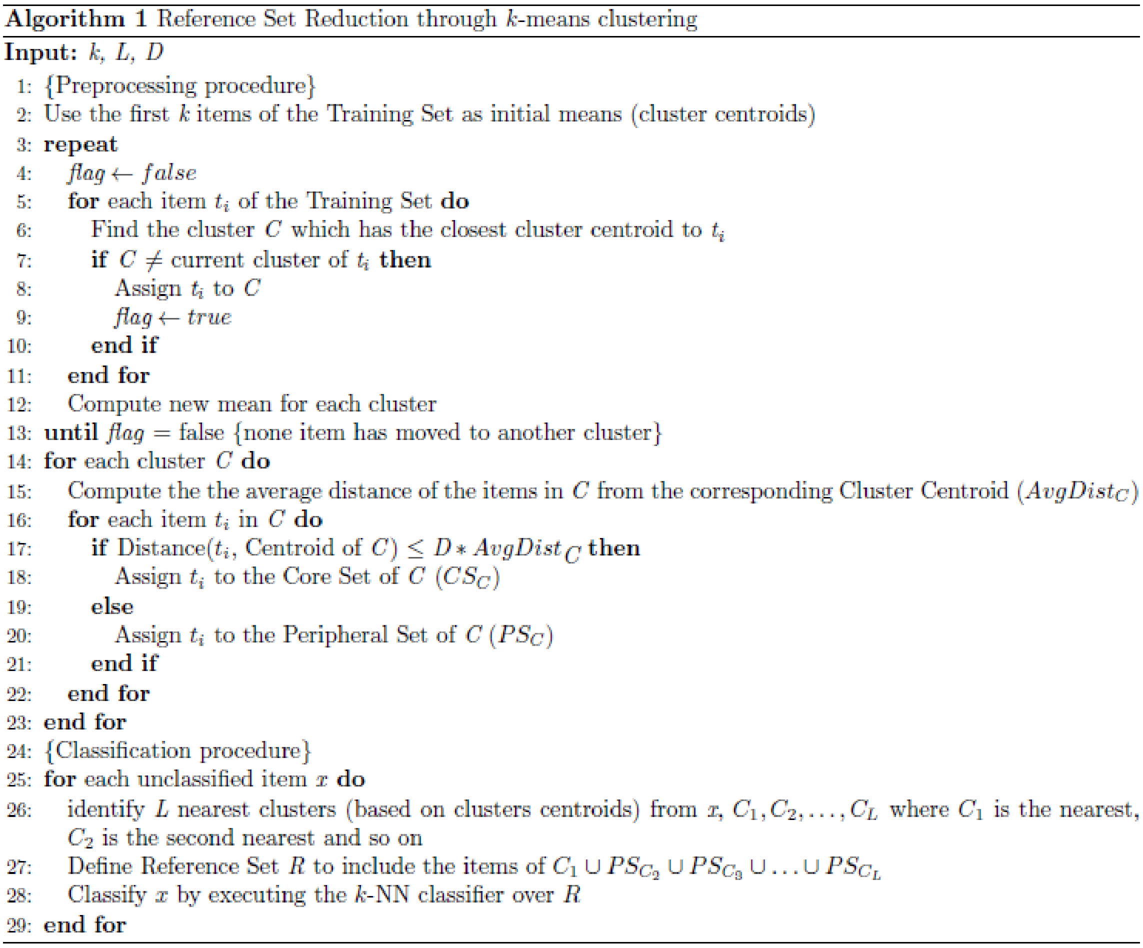
The Reference Set Reduction Method is an adaptive approach which provides three parameters. Its effectiveness depends on the adjustment of these parameters. Hwang and Cho presented experimental results obtained by specific parameter values and based on only one dataset. Moreover, they did not use the well known Euclidean distance as the distance metric. These observations constitute the motivation of our work. Thus, the contribution of this paper is an extensive experimental study on this method. It includes experiments on five real life datasets using different parameter values. We also use as a metric the Euclidean distance.

The rest of this paper is organized as follows. Section II considers in detail the Reference Set Reduction method through k-means clustering and discusses the adaptive schema that it provides. In Section III, we present an extensive experimental study based on five real life datasets. The paper concludes in Section IV.

II. REFERENCE SET REDUCTION THROUGH k-MEANS CLUSTERING

The Reference Set Reduction method (for simplicity, RSRM) proposed by Hwang and Cho is an effective speed-up approach. The method is outlined in Algorithm 1. It uses the well-known k-means algorithm (McQueen, 1967) to find clusters in the training set (lines 2–13). Afterwords, each one cluster is divided into two sets which are called “peripheral set” and “core set”. Particularly, the cluster items lying within a certain distance from the cluster centroid are placed into the “core set”, while the rest, more distant from the centroid, items are placed

* Stefanos Ougiaroglou is supported by a scholarship from the State Scholarship Foundation of Greece (I.K.Y.).



Algorithm 1 Reference Set Reduction through k-means clustering

```
Input: k, L, D
 1: {Preprocessing procedure}
 2: Use the first k items of the Training Set as initial means (cluster centroids)
 3: repeat
 4:   $flag \leftarrow false$
 5:   for each item $t_i$ of the Training Set do
 6:     Find the cluster $C$ which has the closest cluster centroid to $t_i$
 7:     if $C \neq$ current cluster of $t_i$ then
 8:       Assign $t_i$ to $C$
 9:       $flag \leftarrow true$
10:     end if
11:   end for
12:   Compute new mean for each cluster
13: until $flag$ = false {none item has moved to another cluster}
14: for each cluster $C$ do
15:   Compute the the average distance of the items in $C$ from the corresponding Cluster Centroid ($AvgDist_C$)
16:   for each item $t_i$ in $C$ do
17:     if Distance($t_i$, Centroid of $C$) $\leq D * AvgDist_C$ then
18:       Assign $t_i$ to the Core Set of $C$ ($CS_C$)
19:     else
20:       Assign $t_i$ to the Peripheral Set of $C$ ($PS_C$)
21:     end if
22:   end for
23: end for
24: {Classification procedure}
25: for each unclassified item $x$ do
26:   identify $L$ nearest clusters (based on clusters centroids) from $x$, $C_1, C_2, \ldots, C_L$ where $C_1$ is the nearest,
      $C_2$ is the second nearest and so on
27:   Define Reference Set $R$ to include the items of $C_1 \cup PS_{C_2} \cup PS_{C_3} \cup \ldots \cup PS_{C_L}$
28:   Classify $x$ by executing the $k$-NN classifier over $R$
29: end for
```

into the "peripheral set" (lines 14–23).

When a new item x must be classified, the algorithm finds the nearest cluster C1. If x lies within the core area of C1, it is classified by retrieving its k-nearest neighbors from C1. Otherwise, the k nearest neighbors are retrieved from the Reference Set R formed by the items of the nearest cluster and the "peripheral" items of the L most adjacent clusters (lines 25–29).

If the clusters were not divided and only the items of the nearest cluster were used to classify the new item (regardless of how distant from the centroid it was), many training items in the nearby clusters would be ignored. Thus, Hwang and Cho proposed the use of some nearby clusters as a safer approach. The main innovation in their method is that it uses only the peripheral items of these additional adjacent clusters. If all items (not only the peripheral) of these clusters were used, the computational cost would have been much higher.

A key factor of RSRM is the determination of the threshold that defines which items will be core and which peripheral. This is very critical since it determines how many items are accessed during classification. Hwang and Cho consider as peripheral items, those whose distance from the cluster centroid is greater than the double average distance among the items of each cluster. Thus, the average distance among the items in each cluster and the corresponding cluster centroid must be computed (line 15).

In this study, we do not use a particular threshold as Hwang and Cho did (they used the double average distance). We introduce parameter D to be responsible for the splitting of the clusters into core and peripheral sets. An item x is placed into the peripheral set of cluster C, if:

$$\text{Distance}(x, \text{centroid of C}) > D * \text{AvgDistC}$$

For example, if D=1.5, the "peripheral sets" include items that are more than 1.5 times the average distance away from the cluster centroid. The determination of D is a critical issue and it should be made by considering the available number of clusters and the desirable tradeoff between accuracy and computational cost.

Another issue that must be addressed is related to the number of clusters that are constructed (determination of the k parameter in k-means algorithm) and the number of adjacent clusters that are examined when the new item lies within the peripheral area of the nearest cluster (L parameter). Hwang and Cho

Table 1: Dataset description (cost is in million distance computations)

dataset	train/test dataset size	attributes	classes	best k	accuracy (%)	cost
Letter recognition	15000/5000	16	26	4	95.68	75
Magic gamma telescope	14000/5020	10	2	12	81.39	70.28
Pendigits	7494/3498	16	10	4	97.89	26.21
Landsat sattelite	4435/2000	36	6	4	90.75	8.87
Shuttle	43500/14500	9	7	2	99.88	630.75

empirically define L = ⌊ k⌋.

III. EXPERIMENTAL STUDY

The extensive experimental study was conducted using five real life datasets distributed by the UCI Machine Learning Repository[1] . The datasets are presented in Table 1. The fifth column lists the k value found to achieve the highest accuracy when using the k-NN classifier over the whole training set (conv-k-NN). The computational cost was estimated by counting the distance computations needed to carry out the whole classification task. Of course, the cost measurements do not include the distance computations needed by the k-means clustering preprocessing procedure. Moreover, contrary to Hwang and Cho, who used the ROC distance metric in their experiment, we estimated all distances using the Euclidean distance. All datasets were used without data normalization or any other transformation. Also, in all RSRM experiments, we chose the k values of the k-NN classifier that achieved highest accuracy (do not confuse this parameter with k of k-means clustering).

We define L = ⌊√k⌋ as Hwang and Cho did in their experiment. Concerning the k parameter that determines the number of clusters that are formed, we built 8 classifiers for each dataset. Classifier$_i$ uses k = ⌊√n/2i ⌋ clusters, i=1,. . . ,8, where n is the number of items in the training set. Classifier$_1$ is based on the rule of thumb that defines k = ⌊√n/2⌋ (Mardia et al, 1979). We decided to build classifiers that use low k values based on the observation that Hwang and Cho set k=10 for a training set with 60919 items. For each classifier, we chose a varying value for D (1, 1.5, and 2). Thus, we built and evaluated 8 * 3 = 24 classifiers for each dataset.

In Fig. 1–5, for each dataset, the performance of the most accurate classifiers for a given cost are presented[2] . The figures do not include the performance of conv-k-NN that is mentioned in Table 1. In particular, in Fig. 1–5, the classifiers built by the three D values (1, 1.5 and 2) are compared to each other.

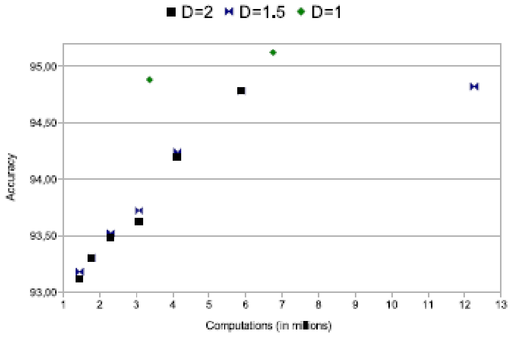


Figure 1: Letter Image Recognition Dataset

Figure 1: Letter Image Recognition Dataset

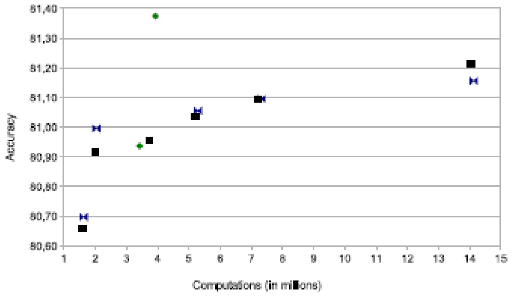


Figure 2: Magic Gamma Telescope Dataset

For the first three datasets (Fig. 1–3), the classifiers built for D=1 seem to perform better than the ones built for D=1.5 and D=2. In the cases of the Letter Image Recognition (LIR) and Magic Gamma Telescope (MGT) datasets, the superiority of the Classifiers D=1 is obvious. In the case of LIR, the two Classifiers D=1 presented in Fig 1 are build by setting $k=\lfloor\sqrt{15000}\ /\ 2_1\rfloor=86$, $L=\lfloor\sqrt{86}\rfloor=9$ and $k=\lfloor\sqrt{15000}\ /\ 2_5\rfloor=21$, $L=\lfloor\sqrt{21}\rfloor=4$, respectively. In MGT, the parameter values of the most accurate classifier are D=1, k=59 and L=7. Finally, in Pendigids, the fastest and slowest Classifier$_{D=1}$ is built by setting k=61 and k=15 respectively.

For the Landsat Satellite (LS) and Shuttle datasets (Fig. 4 – 5) there is not a dominant D parameter value in terms of performance and accuracy. In LS, the most

[1] http://archive.ics.uci.edu/ml/

[2] Detailed experimental results available at:http://users.uom.gr/~stoug/RSRM.zip

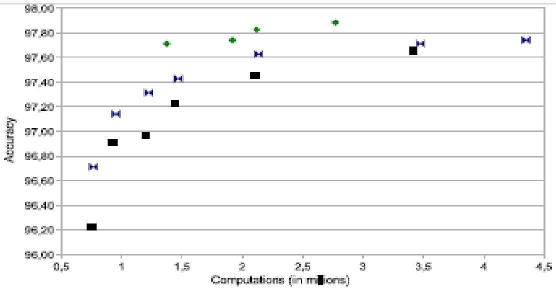


Figure 3: Pendigits Dataset

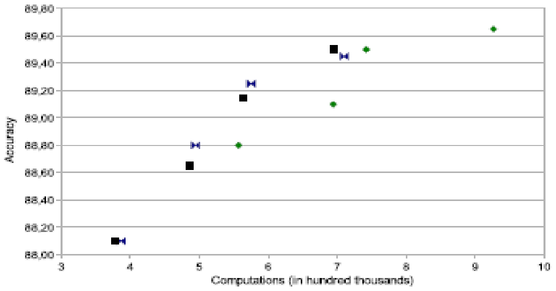


Figure 4: Landsat Satellite Dataset

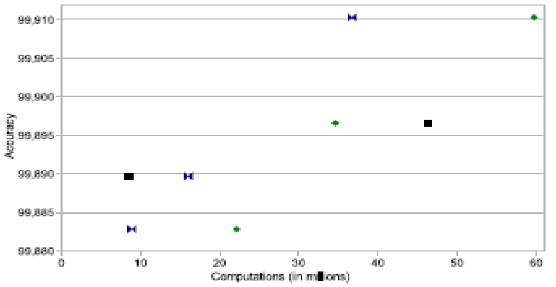


Figure 5: Shuttle Dataset

accurate classifier is built by setting D=1 and k=16, while the fastest classifier that achieves an accuracy value over 89.2% is built using D=1.5 and k=23. In Shuttle, the results are more confusing. This is because Shuttle is an imbalanced (skewed) dataset (approximately 80% of the items belong to one class). However, in Shuttle, all classifiers presented in Fig. 5 manage to achieve higher accuracy than that of the conv-k-NN.

IV. CONCLUSION

In this paper we presented an extensive experimental study on the Reference Set Reduction method through k-means clustering. In all experiments, the well-known Eucledian distance was used. The classification performance of RSRM depends on the determination of k and D parameters. In all cases, they should be adjusted by taking into consideration the application domain and the desirable trade-off between classification accuracy and computational cost. The experimental measurements indicate that if accuracy is more critical than cost, low D and high k and L values (e.g. D=1) lead to an efficient classification method. On the other hand, if cost is more critical than accuracy, higher D and lower k and L values may be more appropriate.

REFERENCES

Dasarathy B. V.,Nearest neighbor (nn) norms: nn pattern classification techniques, IEEE CS Press (1991).

Eick, F. Christoph, Nidal Zeidat and Ricardo Vilalta, Using Representative-Based Clustering for Nearest Neighbor Dataset Editing, in Proc. Fourth IEEE International Conference on Data Mining (ICDM), Brighton, England, 375-378 (2004).

Han, J. and M. Kamber, Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann (2011).

Hwang, S. and S. Cho, Clustering-based reference set reduction for k-nearest neighbor, In Proceedings of the 4th international symposium on Neural Networks (ISSN): Part II–Advances in Neural Networks, Nanjing, China, 880888 (2007).

Lozano M. T. Data Reduction Techniques in Classification processes. Phd Thesis, Universitat Jaume I (2007).

Mardia, K., J. Kent, and J. Bibby. Multivariate Analysis. Academic Press (1979).

McQueen J. Some methods for classification and analysis of multivariate observations, In Proc. of 5th Berkeley Symp. on Math. Statistics and Probability, Berkeley, CA : University of California Press, 281298 (1967).

Samet H., Foundations of Multidimensional and Metric Data Structures. The Morgan Kaufmann Series in Computer Graphics, Morgan Kaufmann Publishers, San Francisco, USA (2006).

Wilson D. R. and T. R. Martinez, Reduction techniques for instance-based learning algorithms, Machine Learning, Vol.:38, 257-286 (2000).

Zezula, P., G. Amato, V. Dohnal, M. Batko, Similarity Search: The Metric Space Approach. Advances in Database Systems, vol. 32, Springer (2006).

Zhang, B. and S. N. Srihari. Fast k-nearest neighbor classification using cluster-based trees, IEEE Transactions Pattern Analysis and Machine Intelligence, vol. 26, 525-528, (2004).