Krassimir Markov, Vladimir Ryazanov,
Vitalii Velychko, Levon Aslanyan
(editors)

New Trends in Classification and Data Mining

I T H E A
SOFIA
2010

Krassimir Markov, Vladimir Ryazanov, Vitalii Velychko, Levon Aslanyan (ed.)

New Trends in Classification and Data Mining

ITHEA®

Sofia, Bulgaria, 2010

First edition

Recommended for publication by The Scientific Concil of the Institute of Information Theories and Applications FOI ITHEA

This book maintains articles on actual problems of classification, data mining and forecasting as well as natural language processing:

- new approaches, models, algorithms and methods for classification, forecasting and clusterisation. Classification of non complete and noise data;
- discrete optimization in logic recognition algorithms construction, complexity, asymptotically optimal algorithms, mixed-integer problem of minimization of empirical risk, multi-objective linear integer programming problems;
- questions of complexity of some discrete optimization tasks and corresponding tasks of data analysis and pattern recognition;
- the algebraic approach for pattern recognition - problems of correct classification algorithms construction, logical correctors and resolvability of challenges of classification, construction of optimum algebraic correctors over sets of algorithms of computation of estimations, conditions of correct algorithms existence;
- regressions, restoring of dependences according to training sampling, parametrical approach for piecewise linear dependences restoration, and nonparametric regressions based on collective solution on set of tasks of recognition;
- multi-agent systems in knowledge discovery, collective evolutionary systems, advantages and disadvantages of synthetic data mining methods, intelligent search agent model realizing information extraction on ontological model of data mining methods;
- methods of search of logic regularities sets of classes and extraction of optimal subsets, construction of convex combination of associated predictors that minimizes mean error;
- algorithmic constructions in a model of recognizing the nearest neighbors in binary data sets, discrete isoperimetry problem solutions, logic-combinatorial scheme in high-throughput gene expression data;
- researches in area of neural network classifiers, and applications in finance field;
- text mining, automatic classification of scientific papers, information extraction from natural language texts, semantic text analysis, natural language processing.

It is represented that book articles will be interesting as experts in the field of classifying, data mining and forecasting, and to practical users from medicine, sociology, economy, chemistry, biology, and other areas.

General Sponsor: Consortium FOI Bulgaria (www.foibg.com).

Printed in Bulgaria

ISBN 978-954-16-0042-9

C\o Jusautor, Sofia, 2010

COMPOSITE BLOCK OPTIMIZED CLASSIFICATION DATA STRUCTURES

Levon Aslanyan, Hasmik Sahakyan

Abstract: *There are different applications that require finding in a set all points most similar to the given query element. These are problems known as classification problems, or as best match, or the nearest neighbor search procedures. The main goal of this paper is to analyze the algorithmic constructions that appear in a model of recognizing the nearest neighbors in binary data sets. Our target is an earlier proposed algorithm [W,1971, R 1974, A,2000] where existence and effectiveness of structures that serve the search procedure is raised. These structures split the search area into the perfect codes and/or the discrete isoperimetry problem solutions, which coexist for a very limited area of parameters. We extend indirectly these parameters through the composite block constructions and estimate the resulting loss of effectiveness of the procedure under consideration.*

Keywords: *Pattern Recognition, Classification, Best Match, Perfect Code*

ACM Classification Keywords*: 1.5. Pattern recognition, H.2.8 Database applications, Data mining*

Introduction

Consider an application scenario.

Let an orthography dictionary and a text file for correction is given. Word by word correction scenario fulfills consecutive comparison of text words with words that are in dictionary. Suppose that a formalism for simple word 'mistake formats' is given – such as one or more wrong letters, single character transpositions, retyping mistakes, missing characters, etc. More detailed formalisms can be devised involving grammatical rules and relations but this simple model quiet well demonstrates the practical problem we consider. Suppose that we are able to define an appropriate measure (metric) between the words (correct and misspelled) – words of text and words of dictionary. Then, it is worthwhile to seek the correction word for the word from text among the closest words of the dictionary by the given metric [GM,2005].

Dramatically this simple scenario speaks about deep similarities between three very different research areas – coding for error correction theory, supervised classification and pattern recognition, and finally – the search for similarities - best matches and nearest neighbors. We do not have in our problem a single bit change like in basic coding theory but we have a larger change through the spelling error model. In terms of pattern recognition we work with too many classes – one for each dictionary word and the learning set (dictionary) is very large. Finally we deal with repeated searches in a single file (dictionary) which obviously is to be well structured to minimize the search time for these specific queries of word spelling. The problem is treating as is, without separating and allocating it to one of the mentioned research areas. Coding is used as the basic structuring tool, compactness hypothesis from pattern recognition is proving the optimality of structures, and search algorithm is composite to achieve the tradeoff between the structural validity and the functional optimality.

The initial structures and investigation is by earlier works of P. Elias and R. Rivest [W,1971, R,1974]. A special cellular block partitioning of basic word space is considered, and a special dynamic programming style mechanism of search of best match or nearest neighbor word sets is applied. Consecutively, there appeared in area different alternative search strategies such as: k-d trees [F,1975], vp-trees [Y,1993], Voronoi tessellation [L,1982], etc. Although new forms and approaches are more perspective, we consider the basic model [W,1971, R,1974] and our additional input is in applying our detailed study and results on discrete isoperimetry problem as

the known formalism for pattern recognition compactness hypothesis [A,1989, A,1981]. Then we construct composite blocks which split the basic set into the homogeneous blocks. Blocks are not isoperimetric so we estimate the resulting loss of search optimality.

There are several typical applications by the same scenario. A data file, for example [F,1975], might contain information on all cities with post offices in a region. Associated with each city is its longitude and latitude. If a letter is addressed to a town without a post office, the closest town that has a post office might be chosen as the destination. The solution to such problem is of use in many other applications too. Information retrieval might involve searching in a catalogue for those items most similar to a given query item; each item in the file would be catalogued by numerical attributes that describe its characteristics. Classification decisions can be made by selecting prototype features from each category and finding which of these prototypes is closest to the record to be classified. Multivariate density estimation can be performed by calculating the volume about a given point containing the closest neighbors.

Basic Structures and Definitions

Let F be a finite set of some binary words of length n. x is an input binary n-word. $F(x)$ denotes the set of all words from F having the (same) minimal possible distance from x (in the simplest case the Hamming distance is applied). Optimization and complexity issues of algorithms, working with F and x and composition of $F(x)$ is considered, concerning with special constructions that map the basic set F onto the computer memory. The initial methods, based on error correcting perfect codes are defined in [W,1971, R,1974]. They are restricted by the very limited set of possible perfect codes that limits the special constructions mentioned above. It is well known that the only nontrivial classes of binary perfect codes are Hamming and Golay codes [TP,1971, ZL,1972]. We use the geometric interpretation, when the code centered and none intersecting sphere systems are given. Linear codes allow optimizing of addressing issues in considered models. Spheres play the role of blocks when overall algorithm optimality requires the discrete isoperimetry property (DIP) of blocks.

DIP problem is one of the typical issues of advanced discrete mathematics. DIP solutions are very close to the Hamming spheres geometrically. The complete coverage of basic binary set by such objects is highly important and provides more possibilities of construction of searching algorithms. Any positive steps on this direction require more knowledge on properties of the solutions of the DIP that are provided by authors in [A,1989 - A,2000].

So, let us suppose that the basic set Ξ^n of all binary words of length n is divided into the blocks $B_1, B_2, \ldots, B_m$. Accordingly, $L_i = F \cap B_i,\ i = 1,2,\ldots,m$ are the lists of elements of F belonging to these blocks. They are saved as separate lists by addresses $h(B_i)$, common for all elements of each such block of Ξ^n. The idea of this searching construction is transparent - using a dynamic programming style algorithm of class of branches and leaves and a partitioning the space Ξ^n into the geometrically compact blocks, the algorithm may apply to less input information to get the $F(x)$ by an input x. To achieve this result we have to divide Ξ^n into the disjoint blocks - solutions of the DIP [R,1974]. For example, splitting into the spheres by a perfect code is ideal. And the problem is that neither the splitting of Ξ^n into the solutions of the isoperimetry problem nor the construction of a perfect code for an arbitrary n is possible. Alternatives are to be constructed and then evaluated.

Let us summarize the main points of structuring Ξ^n for search:

Ξ^n is divided into the simple blocks $B_1, B_2, \ldots, B_m$. Blocks are similar in their sizes, and the lists of elements in blocks $L_i = F \cap B_i,\ i = 1,2,\ldots,m$ practically have comparable sizes.

Given an arbitrary x, blocks can be quickly and simply ranged by their distances to x.

The simplest partition as we mentioned is by a Hamming code. But these codes exist for dimensions $n = 2^q - 1$ only. And the block size by Hamming codes is very much limited: $n+1$. The same time ranging blocks by the input vector x is very quick and ideally simple. Other codes such as Golay codes and non linear codes or codes in other distances do minimal help. It is also to take into account the quasi-perfect and nearly perfect codes which provide more constructions for approximating the problem. Next issue is the search optimality that depends on block shapes. The main point of optimality is proposed in [R,1974]:

The block is optimal when its shape is the DIP solution.

Let us comment the proposed optimality of DIP solutions in [R,1974]. In a supposition that F is a random set of vertices of Ξ^n with membership probability p, probabilities of blocks that are analyzed for $F(x)$ are evaluated. The following resolution is used:

$$\Psi(C) - \Psi(B) = 2^{-n}\left(\left|C^{(m)}\right| - \left|B^{(m)}\right|\right)\Theta(n, m-1) + \tag{1}$$

$$2^{-n}\sum_{m<\delta\le n}\left(\left|C^{(\delta)}\right| - \left|B^{(\delta)}\right|\right)\Theta(n, \delta-1) \ge \tag{2}$$

$$2^{-n}\left(\left|C^{(m)}\right| - \left|B^{(m)}\right|\right)(\Theta(n, m-1) - \Theta(n, m)) \ge 0 \tag{3}$$

Here C and B are blocks of the same size, and B is DIP optimal. In fact B is spherical [R,1974] or initial segment of standard placement L_n (see below) [A,1989]. $B^{(i)}$ is the i neighborhood of B - the vertices that are in distance i from B. $\Theta(n,m)$ is probability that a sphere of radius m is empty (of F). m is the position/index, that $\left|C^{(i)}\right| = \left|B^{(i)}\right|$ for $i = 0,\ldots,m-1$ (that probably may happen) but $\left|C^{(m)}\right| > \left|B^{(m)}\right|$. Because of DIP optimality of B:

$\left|C^{(i)}\right| > \left|B^{(i)}\right|$ for some number of indexes i after m, then this may become negative, and in case $\left|C^{(i)}\right|$ (4) may become 0.

[R,1974] formulates the lexicographic minimality of $\left|B^{(0)}\right|,\ldots,\left|B^{(i)}\right|,\ldots,\left|B^{(k)}\right|$. This is not satisfactory for transfer from (2) to (3). The stronger and satisfactory is (4) that we brought above in accord to DIP postulations of [A,1989].

Consider a set $A \subseteq \Xi^n$. A point $\alpha \in A$ is called the inner point of A if the unit sphere $S_n(\alpha) \equiv \{\beta \in A \mid \rho(\alpha,\delta) \le 1\}$ at α, is a member of A. In opposite case we call α the boundary point of A. Let $\Im(A) \subseteq A$ is the collection of all inner points of A. Then $\Re(A) = A \setminus \Im(A)$ is the set of all boundary vertices. A set $A \subseteq \Xi^n$ is called DIP optimal (isoperimetric) if $\left|\Im(A)\right| \ge \left|\Im(B)\right|$ for any $B \subseteq \Xi^n$, $\left|B\right| = \left|A\right|$.

Consider the linear order L_n of vertices Ξ^n called standard. L_n is the order of vertex sets of layers of Ξ^n from 0 to n, and inside the layers - order is lexicographical. Let a is a nonnegative integer, $a \le 2^n$. Denote by $L_n(a)$ the initial a-segment of L_n, then the Main Isoperimetric Theorem proves that $L_n(a)$ is a solution of DIP [A,1989]. This result was formulated independently by different authors.

Call the number m k,δ-spherical, if $m=\sum_{i=0}^{k}C_n^i+\delta$. As it has been proposed in [R,1974] the optimal best match search algorithms of hashing class correspond to the selection of blocks B_i which are the DIP solutions. To apply this result one need to split Ξ^n into such blocks. We have:

The arbitrary solution of the DIP contains a Hamming sphere of the maximal possible sphere by the given m (the radius of these sphere equals k) and only the additional δ vertices or the part of these might be arranged differently.

At least for 2^{n-1} cases by all different m the more precise description of solutions of DIP is given. These DIP solutions are included in a sphere of radius $k+2$. Such numbers m we call critical. The quantity of vertices between some two closest critical numbers is distributed randomly and they can't create additional inner vertices.

The number of subsets of Ξ^n with ξ interior vertices is distributed by the Poisson's distribution with the main value ½. This is in case of random membership of vertices of Ξ^n into the considered sets by the probability ½. For probabilities other than ½ the picture is similar but more complicated, which is described in a separate paper.

All the above mentioned descriptions of DIP solutions are rather complicated to construct the precise splitting of Ξ^n by these objects. So the geometrical shape of the DIP solutions give us the ideal partitioning objects exemplifying.

Coming back to the formula (3) let i_0 be the minimal index i with $\left|C^{(i)}\right|=0$. We take $\Theta(n,i_0)$ instead of $\Theta(n,m)$ in (3) for indexes i_0 and above. For parts where $\left|C^{(i)}\right|-\left|B^{(i)}\right|$ are positive and negative we may take some evaluating values $\Theta(n,i_1)$ and $\Theta(n,i_2)$ which is also possible. Then it is easy to see that $\Theta(n,i_1)$ satisfies (3) because of $\sum_{m\le\delta\le n}\left|C^{(\delta)}\right|=\sum_{m\le\delta\le n}\left|B^{(\delta)}\right|$.

Recall the main requirements to block structures: partitioning of Ξ^n; computation of distances; and DIP optimality. The Hamming sphere is the simplest DIP solution. For other block sizes DIP solutions are similar but different. First question arises is about the approximate block partitioning of Ξ^n. One approach is in partitioning into the DIP solutions with intersections. If diversity of DIP solutions is high and/or intersections can be minimized then the optimality loss is related to repetitions of elements in different lists which may be small. The second approach is in splitting the space into the subspaces for which block partitioning is effective. Even small blocks in subspaces become large in a Cartesian product. The optimality loss is related to non DIP optimality of product which is estimated below. Before that we mention in short some similar structures from the coding theory. Here accent is exactly on reduction of repetitions of elements in intersections and not to the DIP optimality.

An (n,t)-quasi-perfect code is a code for which the spheres of radius t around the code words are disjoint, and every vector is at most $t+1$ from some code word. A subclass of quasi-perfect codes is nearly perfect codes. A few nearly perfect code sets are known. For $t=1$ there exists a nearly perfect code for any $n=2^q-2$. For $t=2$ there exist a nearly perfect code for $n=4^q-1$. It is also known that no other nearly perfect codes exist.

For any nearly perfect code: vectors with a greater than t distance from any code word is at distance $t+1$ from exactly $\lfloor n/(t+1)\rfloor$ other code words; and vectors of distance t from some code word are at distance $t+1$ from exactly $\lfloor (n-t)/(t+1)\rfloor$ other code words.

Composite Block Structures

Second approach concerned to effective structuring for best match search is to analyze the possibility of structuring the so called composite blocks. This is when we split the basic space Ξ^n into the Cartesian product of smaller size spaces. Then we use different constructions, based on the Hamming spheres of different sizes. Given an n we first choose a number of form $2^q - 1$ to be not greater than n. As we know there exists a perfect Hamming code for this case, so, the exact partitioning of Ξ^n into the spheres of radius 1 is given and may be used. Considering partitioning of Ξ^n according with the Cartesian product of Ξ^{2^q-1} and Ξ^{n-2^q+1}, we can choose the splitting of the first subspace as a Hamming code while the second part might be reminded as is or further split by itself. The main advantage is that the blocks are Hamming spheres and that the values of corresponding $h(x)$ functions as well as the distances of these blocks from the arbitrary points $x \in \Xi^n$ are simply computable.

Generalization of this idea is related to the special representation of arbitrary numbers n by the sums of numbers of form $2^q - 1$. This is for decomposition of Ξ^n into the subcubes, which can be covered by the sets of disjoint Hamming spheres. In parallel additional compounds may be used such as Golay codes and the simple constructions which partition arbitrary cubes into the two subspheres, etc.

To be compact we can formulate the following properties:

Let us consider a binary vector $\tilde{\alpha} = (\alpha_n, \alpha_{n-1}, ..., \alpha_1)$. The corresponding sum $s(\tilde{\alpha}) = \sum_{i=1}^{n} \alpha_i (2^{i-1} - 1)$ is limited by numbers 0 and $2^{n-1} - n$. Moreover, all sums of form $s(\tilde{\alpha})$ don't cover this interval completely.

Sums $s(\tilde{\alpha})$ achieve the 2^{n-1} different values. The last coordinate - α_1 is not essential for the value of $s(\tilde{\alpha})$.

Let us consider the vector $\tilde{1}_i$ with the all 1 coordinates on positions $j, j \geq i$ and 0 elsewhere. We'll double the last sum term - $2^{i-1} - 1$, which corresponds to the i-th 1 of $\tilde{1}_i$. Then we get the numbers from $2^n - 2$ to $2^n - n$ continuously. For the numbers, starting from $2^n - n - 1$ and smaller we can prove by induction on n, that doubling only one sum term we can receive an arbitrary number from the remainder part.

Consider a simple case of composite blocks which are in 2 parts. Let $\Xi^n = \Xi^{2^{q_1}-1} \times \Xi^{2^{q_2}-1}$. Consider Hamming codes in $\Xi^{2^{q_1}-1}$ and in $\Xi^{2^{q_2}-1}$ and define the blocks as Cartesian product of unit spheres defined by these codes. Such block is included in a sphere of radius 2. Similarly in the case of Cartesian product of two arbitrary spheres – one of radius k_1 and second of radius k_2, the result is included in sphere of radius $k_1 + k_2$. The points of these spheres that are not a block member are in different layers. In second layer for example there are $k_1 \cdot k_2$ points. This formula can be extended for other layers but we prefer to compute the missing points from another point of view.

First claim is that constructed blocks consist of $b = \sum_{i_1=0}^{k_1} C^{i_1}_{2^{q_1}-1} \cdot \sum_{i_2=0}^{k_2} C^{i_2}_{2^{q_2}-1}$ points each. Number of points of $k_1 + k_2$ sphere is $s = \sum_{i=0}^{k_1+k_2} C^{i}_{2^{q_1}+2^{q_2}-2}$. The number of missing point of block to the complete sphere equals $s - b$. Compare s and b. In a simplified structure when $q_1 = q_2 = q$ and $k_1 = k_2 = k$ we apply the following known inequality [PW,1972]:

$C_n^{\lambda n} < \sum_{i=0}^{\lambda n} C_n^i < \frac{1-\lambda}{1-2\lambda} C_n^{\lambda n}$, $\lambda < 1/2$.

Here λn supposed to take integer values. Blocks in general are not very large so that we suppose the sphere radius is some constant number. Let $\lambda = o(1)$ with $n \to \infty$. Then $b \sim (C_{2^q-1}^k)^2$ and $s \sim C_{2(2^q-1)}^{2k}$. For small k, the Hamming case included, these values are comparable.

Conclusion

Finding nearest neighbors is a regular procedure in experimental data analysis. Pattern recognition is the closest model where the nearest elements of the learning set is a question. To quick up the search for similarities it is common to use divide and conquer approach through the partition of unit cube into the blocks. Three requirements are the main: partitioning; computation of distances; and optimality. The tradeoff between these concurring requirements can be resolved partially. DIP solutions and standard placement in particular, perfect and nearly perfect codes that exist for exceptional dimensions, and space partitioning into the Cartesian products are the main algorithmic resource. Complete solutions are linked to perfect codes and all other cases are accompanied with losses and approximations. Future work will describe similar structures in other metrics, e.g. Lee metrics.

Bibliography

[W, 1971] T.A. Welch. Bounds on the information retrieval efficiency of static file structures, Project MAC Rep. MAC-TR-88, Mass. Inst. of Tech., Cambridge, Mass., 164 p., 1971, Ph.D. thesis.

[R, 1974] R.L. Rivest. On The Optimality of Elias's Algorithm for Performing Best-Match Searches, Information Processing 74, Nort-Holland Publishing Company, pp. 678-681, 1974.

[F, 1975] J. H. Friedman, F. Baskett and L.J. Shustek, An algorithm for finding nearest neighbors, IEEE Trans. Comput., vol. C-24, pp. 1001-1006, Oct. 1975.

[Y, 1993] Yianilos, Peter N., Data structures and algorithms for nearest neighbor search in general metric spaces, Proceedings of the fourth annual ACM-SIAM Symposium on Discrete algorithms, Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 311–321, 1993.

[L, 1982] D.T. Lee, On k-Nearest Neighbor Voronoi Diagrams in the Plane, IEEE Transactiona on Computers, Vol. C-31, N. 6, June, pp. 478-487, 1982.

[A, 1989] L.H. Aslanyan, The discrete isoperimetric problem and related extremal problems for discrete spaces, Problemy Kibernetiki, Moscow, v. 36, pp. 85-128, 1989.

[A, 1981] L. Aslanyan and I. Akopova, On the distribution of the number of interior points in subsets of the n-dimensional unit cube, Colloquia Mathematika Societatis Yanos Bolyai, 37, Finite and Infinite Sets, (Eger) Hungary, pp. 47-58, 1981.

[A 2000] L. Aslanyan, Metric decompositions and the Discrete Isoperimetry, IFAC Symposium on Manufacturing, Modelling, Management and Control, July 12-14, Patras, Greece, pp. 433-438, 2000.

[AC, 2007] L. Aslanyan and J. Castellanos, Logic based Pattern Recognition - Ontology content (1), Information Theories and Applications, ISSN 1310-0513, Sofia, Vol. 14, N. 3, pp. 206-210, 2007.

[AR, 2008] L. Aslanyan and V. Ryazanov, Logic Based Pattern Recognition - Ontology Content (2), Information Theories and Applications, ISSN 1310-0513, Sofia, Vol. 15, N. 4, pp. 314-318, 2008.

[TP, 1971] A. Tietäväinen and A. Perko, There are no unknown perfect binary codes, Ann. Univ. Turku, Ser. AI, 148, pp. 3-10, 1971.

[ZL, 1972] V. Zinoviev and V. Leontiev, On the Perfect Codes, Problems of Information Transferring, Moscow, Vol. 8, N. 1, pp. 26-35, 1972.

[PW, 1972] W. W. Peterson and E. J. Weldon, Error-Correcting codes, second edition, The MIT Press, Cambridge, 1972.

[GM, 2005] H. Ghazaryan and K. Margaryan, Armenian spell checker, "Computer Science & Information Technologies" Conference, Yerevan, September 19-23, pp. 632-634, 2005.

Authors' Information

Levon Aslanyan *– Head of Department, Institute for Informatics and Automation Problems, NAS RA, P.Sevak St. 1, Yerevan 14, Armenia, e-mail: lasl@.sci.am*

Hasmik Sahakyan *– Leading Researcher, Institute for Informatics and Automation Problems, NAS RA, P.Sevak St. 1, Yerevan 14, Armenia, e-mail: hasmik@ipia.sci.am*