INFORMED *k*-MEANS: A CLUSTERING PROCESS BIASED BY PRIOR KNOWLEDGE

A case study in the dactyloscopic domain

Wagner Francisco Castilho
Federal Savings Bank and Catholic University of Brasília, Brasília, Brazil

Hércules Antônio do Prado
Brazilian Enterprise for Agricultural Research and Catholic University of Brasília, Brasília, Brazil

Marcelo Ladeira
University of Brasília, Brasília, Brazil

Keywords: Knowledge Discovery in Databases (KDD), Clustering Analysis, Dactyloscopy.

Abstract: Knowledge Discovery in Databases (KDD) is the process by which unknown and useful knowledge and information are extracted, by automatic or semi-automatic methods, from large amounts of data. Along the evolution of Information Technology and the rapid growth in the number and size of databases, the development of methodologies, techniques, and tools for data mining has become a major concern for researchers, and has led, in turn, to the development of applications in a variety of areas of human activity. About 1997, the processes and techniques associated with cluster analysis had begun to be researched with increasing intensity by the KDD community. Within the context of a model intended to support decisions based on cluster analysis, prior knowledge about the data structure and the application domain can be used as important constraints that lead to better results in the clusters' configurations. This paper presents an application of cluster analysis in the area of public safety using a schema that takes into account the burden of prior knowledge acquired from statistical analysis on the data. Such an information was used as a bias for the *k*-means algorithm that was applied to identify the dactyloscopic (fingerprint) profile of criminals in the Brazilian capital, also known as Federal District. These results was then compared with a similar analysis that disregarded the prior knowledge. It is possible to observe that the analysis using prior knowledge generated clusters that are more coherent with the expert knowledge.

1 INTRODUCTION

Fayyad (1996) argues that KDD is the process of extracting new, useful, and interesting knowledge from databases. This process has an iterative and interactive nature and is composed of a series of activities, which includes, as well, previous knowledge and the adequate interpretation of results.

KDD is a field in which various areas related to knowledge converge, integrating mature technologies associated with Statistics, Databases, Machine Learning, Computational Intelligence, Data Warehouse, Artificial Intelligence, and Standards Recognition. Its application has also spread to various areas of human activity, such as finance, science, government, health care, sales and marketing, health insurance and plans, transportation, industry, among others.

This paper presents an application of KDD techniques in the area of public safety, which centers on identifying the patterns corresponding to the dactyloscopic (fingerprint) profile of criminals in the Brazilian capital (Federal District), in comparison to

Francisco Castilho W., Antônio do Prado H. and Ladeira M. (2004).
INFORMED k-MEANS: A CLUSTERING PROCESS BIASED BY PRIOR KNOWLEDGE - A case study in the dactyloscopic domain.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 469-475
DOI: 10.5220/0002646704690475

the national profile, on the basis of the application of a clustering task and statistical resources.

2 INFORMED CLUSTERING

Figure 1 presents a general scheme of the informed clustering process. Departing from the information regarding to the data structure and the application domain the clustering analysis objectives and expectations are defined. Subsequently, it is selected a set of variables or attributes of the objects that are relevant or discriminant within the classification problem under consideration. In constructing an information matrix, we can consider the gradations of interest or relevance of the attributes, as well as the implication and correlation maps within them. Information on the rules of production can influence the homogeneity coefficient and guide the specification of prior knowledge as hypotheses that are introduced at the time the algorithm is reallocated in the search for a better configuration.

3 RECOGNITION OF PATTERNS IN DACTYLOSCOPIC DOMAIN

The purpose of Biometrics is to identify an individual based on his or her physical characteristics. In conjunction with the resources offered by Information Technology, Biometrics offers interesting and effective solutions in the area of public safety, particularly in the identification of individuals involved in criminal activity. Dactyloscopy is a biometric technique that has been widely used to identify criminals, given that it satisfies the requirements of the permanence, immutability, and singularity of fingerprints (Oliveira, 2003). Dactyloscopy is the process by which individuals are identified through the examination of their fingerprints. The digital impression is the mirror image of the digital pattern.

3.1 Dactyloscopic classification system

Dactyloscopic classification systems were developed to reduce the complexity of and time required for the identification of fingerprints. Two main classification systems have been adopted around the world: Vucetich and Henry. The Brazilian police force employs the Vucetich system, the most widely used method in the world. Vucetich defined four primary types of digital impressions in his system with the following classifications: arches, internal loops, external loops, and whorls. Subsequently, the accidental, scar, and amputation types were added. These seven primary types are defined (INI, 1987):

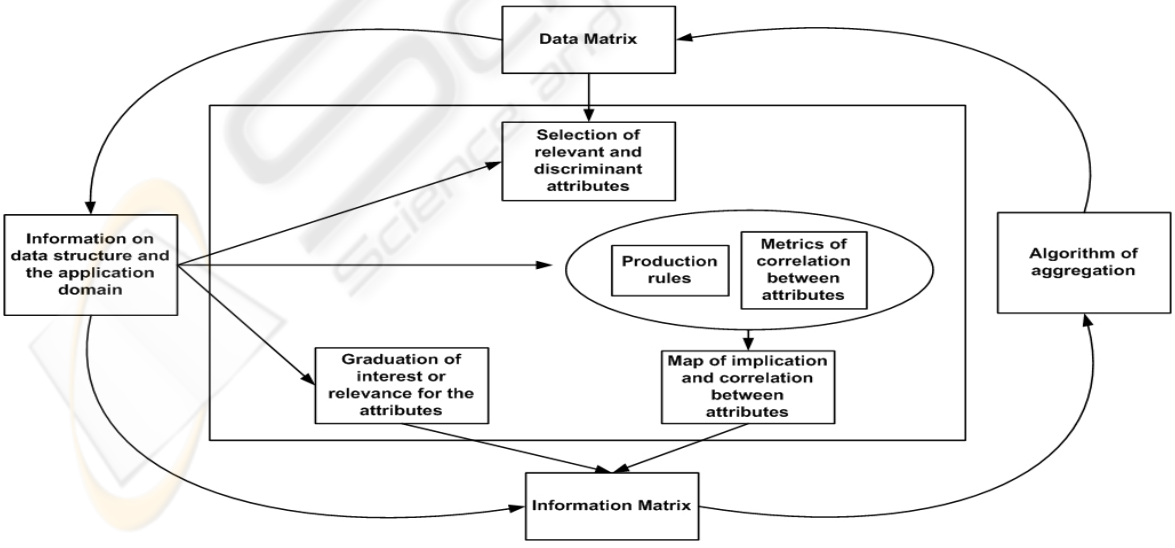


Figure 1: Informed clustering process

Arch - refers to the dactylogram made up of generally parallel and convex ridges that run or tend to run from one side of the print to the other and very often reveal angular or vertical ridges. Represented by the number 1 or the letter A.

Internal Loop: refers to the dactylogram that presents a delta to the observer's right and a nucleus composed of one or more ridges, which run from the left of the print toward the center, recurving and returning, or tending to return, to the side from which they originated, thereby forming one or more loops. Loops involve the two-way movement of a papillary line, which must have perfect inflection. Represented by the number 2 or the letter I.

External Loop: refers to the dactylogram that reveals a delta to the observer's left and a nucleus composed of one or more ridges that run from the left of the print toward the center, recurving and returning, or tending to return, to the side from which they originated, thereby forming one or more loops. Represented by the number 3 or the letter E.

Whorl: refers to the dactylogram characterized by the presence of a delta to the observer's left and right and a varied nucleus, which presents at least one curved ridge in front of each delta. Represented by the number 4 or the letter W.

Accidental: refers to the dactylogram that does not fit within any of the four primary types cited before and which is represented by the number 5.

Scar: refers to the dactylogram that presents a permanent mark caused by a cut, pustule, burn, or crushing, thereby making its classification within one of the 5 types cited above impossible and which is represented by the number 6.

Amputation (or failure): refers to the type in which a total or partial loss of the phalange is evidenced, therefore compromising or even precluding the classification of the primary type, and which is represented by the number 7.

If we create a fraction in which the numerator is the number formed by the numbers that represent the pattern of the fingers of the right hand, extending from the thumb to the small finger, and the denominator constitutes the same number for the left hand, we arrive at the *dactyloscopic formula*, as it is known.

Two fingerprints will only be considered identical when they demonstrate twelve or more characteristic points having the same configuration and location. In the majority of countries, these criteria are required by law for purposes of a positive identification in criminal cases.

4 CASE STUDY

The purpose of the analysis was to identify the pattern of the dactyloscopic (fingerprint) profile of criminals in the Federal District, in comparison to the national profile, on the basis of the application of a clustering analysis and statistics, supported by a clustering model that uses prior knowledge.

It is the task of the National Identification Institute (INI), a branch of the Federal Police Department (DPF), linked to the Ministry of Justice, founded in 1963 and headquartered in Brasilia, to centralize information and fingerprints associated with the subjects of police investigations or individuals charged with crimes within the territorial boundaries of Brazil, as well as foreign nationals subject to registration, through the use of the dactyloscopic identification process. The Dactyloscopic Research Section has an Individual Dactyloscopic Archive (AID) comprised by 19 manual archiving machines for individual dactyloscopic criminals, model NG Class 5500, in which approximately 1,360,000 records are stored. Those records have ten fields in which the ten fingerprints are stored. The archiving of the individual dactyloscopics is initially accomplished on the basis of the fundamental types established in the classification key. The Dactyloscopic Formula (FD) is the set of numerical symbols representing the primary classification of the AID.

The database, known as "MECA-Sinic", was extracted from the DPF's mainframe in November 2000 by a domain expert. The database has a total of 502,052 registries. It represents a sample of 37% of the total number of identification records, randomly extracted. Complete attribute types: criminal violation code, sex, skin, birth date, and main types for each finger. Text attribute types: State. From the database, all the State's records matching those of the Federal District were selected, specifically, a total of 5,363. The attributes selected for the clustering analysis were the 10 primary types corresponding to each finger.

4.1 Incidence of the fundamental dactyloscopic types

According to Araújo (2003), considering the country as a whole, the averages corresponding to the statistics by incidence of the fundamental types are: Whorl (31.16%), Internal Loop (30.84%), External Loop (29.21%), Arch (7.50%). The primary types, Accidental, Scar, and Amputation, combined, do not reach 1.5% of the cases. Of the total, 90.16% were found to be men and 9.84% women. The right hand was found to be predominant in the External Loop type, while the left hand proved predominant in the Internal Loop type, for both sex (the principle of symmetry). Both the right and left index fingers have the highest degree of frequency distribution. The small finger revealed the lowest degree. This highlights the long-standing error in Brazil of using the right thumb instead of the right index finger as the standard for identification documents.

Table 2 presents the percentage of the fundamental types compared by sex. There is a higher incidence among men of the Whorl and Internal Loop types, while for women higher incidences are found in the External Loop and Arch types. The analysis found that among women, the most frequent fundamental type is the External Loop (32.11%), followed by the Internal Loop (29.43%), Whorl (28.01%), and Arch (9.59%), while men displayed a higher incidence of the Whorl type (31.51%), followed by the Internal Loop (30.99%), External Loop (28.89%), and Arch (7.28%).The small fingers have the highest absolute incidence of frequency among the fundamental types, External Loop (83.35%, Right Small Finger) and Internal (80.17%, Left Small Finger) in women. Among men, the inverse is found: Internal Loop (81.80%, Left Small Finger) and External (78.84%, Right Small Finger).

Table 1: Incidence of fundamental types (%)

Hand	Finger	Arch	Internal Loop	External Loop	Whorl
Rigth	Thumb	3,16	0,56	45,36	50,44
	Index	15,42	13,40	35,13	34,02
	Middle	8,54	1,22	69,67	19,06
	Ring	3,05	1,05	47,22	47,72
	Small	2,75	0,38	79,28	16,43
Left	Thumb	5,54	51,88	0,78	41,16
	Index	16,93	36,51	12,78	31,22
	Middle	11,58	66,04	1,10	19,84
	Ring	4,20	55,70	0,55	38,59
	Small	3,85	81,64	0,21	13,14
Average		7,50	30,84	29,21	31,16

Table 2: Fundamental types by sex (%)

Hand	Finger	Arch		Internal Loop		External Loop		Whorl	
		M	W	M	W	M	W	M	W
Right	Thumb	2,93	5,27	0,56	0,56	44,85	49,97	51,16	43,86
	Index	15,38	15,78	13,89	8,83	34,37	42,09	34,24	32,02
	Middle	8,41	9,70	1,29	0,57	69,14	74,51	19,59	14,20
	Ring	2,93	4,14	1,05	0,97	46,43	54,52	48,59	39,70
	Small	2,57	4,40	0,39	0,29	78,84	83,35	17,00	11,15
Left	Thumb	5,27	8,07	52,20	48,98	0,72	1,26	41,15	41,26
	Index	16,61	19,89	36,72	34,60	12,78	12,80	31,25	30,93
	Middle	11,15	15,56	66,29	63,73	1,08	1,31	19,99	18,43
	Ring	3,96	6,41	55,71	55,57	0,51	0,95	38,81	36,52
	Small	3,54	6,67	81,80	80,17	0,19	0,35	13,26	12,04
Average		7,28	9,59	30,99	29,43	28,89	32,11	31,51	28,01

4.2 Incidence of the fundamental dactyloscopic type in the Federal District

The statistics for the data corresponding to the Federal District are presented in table 3. Notice the following statistics of fundamental types: Internal Loop (81.79%), the left small finger; External Loop (80.10), the right small finger; Whorl (50.81%), the right thumb; Arch with an incidence of 16.89% in the left forefinger, followed by 14.64% in the right index finger.

The overall average incidence of the fundamental types in the Federal District reveals only small absolute difference in relation to the national statistics: lower arch (0.17) and internal loop types (0.04); higher external loop (0.19) and whorl types (0.41). Figure 3 presents the different slopes for the Federal District data relative to the national data for each fundamental type and for each finger.

Table 3: Fundamental types in the Federal District (%)

Hand	Finger	Arch	Internal Loop	External Loop	Whorl
Rigth	Thumb (RT)	3,16	0,56	45,36	50,44
	Index (RI)	15,42	13,40	35,13	34,02
	Middle (RM)	8,54	1,22	69,67	19,06
	Ring (RR)	3,05	1,05	47,22	47,72
	Small (RS)	2,75	0,38	79,28	16,43
Left	Thumb (LT)	5,54	51,88	0,78	41,16
	Index (LI)	16,93	36,51	12,78	31,22
	Middle (LM)	11,58	66,04	1,10	19,84
	Ring (LR)	4,20	55,70	0,55	38,59
	Small (LS)	3,85	81,64	0,21	13,14
Average		7,50	30,84	29,21	31,16

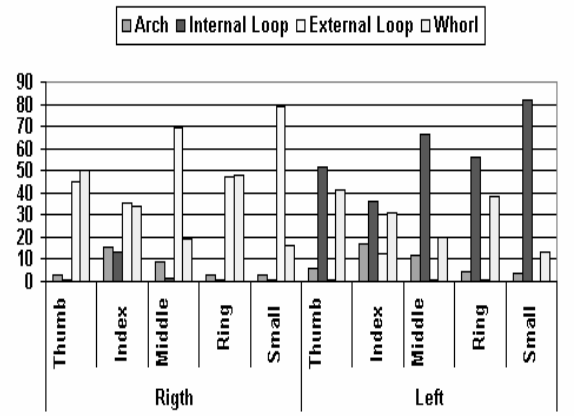
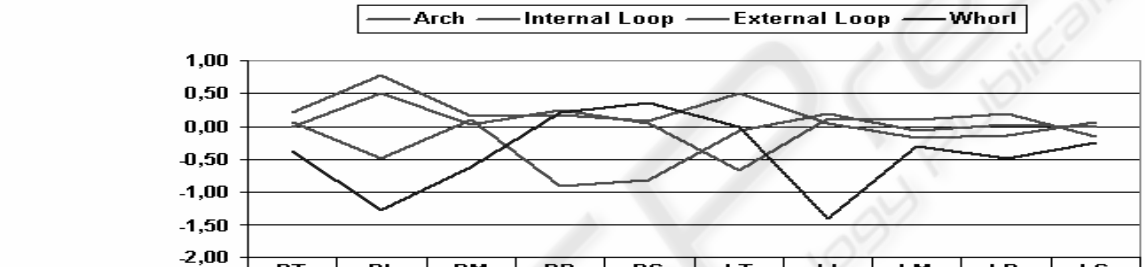


Figure 2: Fundamental types in the Federal District (%)

Experiments with clustering in a number *k* of different clusters restrict the decision for 6, 5, or 4 groups. Experiments with clustering in a number *k* of 6 and 5 groups demonstrated the proximity of the third and fifth groups produced. Thus, in light of this particular configuration for the data structure and the domain application, and following numerous trials using samples, a total number of four clusters was chosen for purposes of the segmentation of the database containing 7,953 registries, corresponding to the total volume of data for the Federal District.



	RT	RI	RM	RR	RS	LT	LI	LM	LR	LS
Arch	0,22	0,78	0,17	0,17	0,08	0,51	0,04	-0,18	-0,13	0,07
Internal Loop	-0,01	0,50	0,03	0,26	0,04	-0,67	0,12	0,10	0,19	-0,15
External Loop	0,07	-0,48	0,11	-0,90	-0,82	-0,06	0,19	-0,07	0,02	0,00
Whorl	-0,37	-1,27	-0,61	0,22	0,36	-0,01	-1,41	-0,30	-0,48	-0,24

Figure 3: Difference between the slopes of the fundamental types in Brasil and the Federal District

4.3 Distribution of the dactyloscopic formulas

In Vucetich's dactyloscopic classification system, 7^{10} alternative formulas can occur, all constructed on the basis of the combination of the seven primary types and referring to each finger on both hands. In the database originally constructed by the domain experts, containing 502,052 registries, an occurrence of only 36,175 formulas was verified. Of the total number of dactyloscopic formulas verified in the original database, five formulas were found to have higher frequencies, representing 11.26% of the occurrences. The dactyloscopic formulas with the highest incidences are set out in figure 4. The remaining occurrences of less frequent formulas (88.74%) are mostly spread out in a distribution of individual frequency (90.72%) of less than 1% (Araújo, 2003).

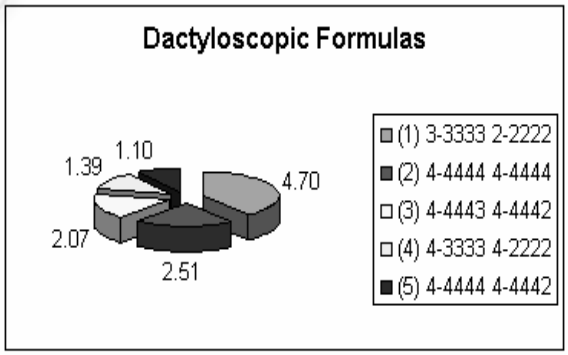


Figure 4: Dactyloscopic formulas with highest incidences (%)

4.4 Results of the clustering

Tables 4 and 5 present the comparative results of the sum of the residual square (*SQRes*) and the information on the configuration of the resulting clusters for the standard algorithm and the classification process supported by prior knowledge.

The results presented indicate that the informed clustering process performed better. Within the process, the *SQRes* was smaller and the dactyloscopic formulas were more evenly distributed among the four groups, with formulas 3 and 5 located within the same group. In cluster 1, 3,340 objects were allocated with a greater resemblance to formula 1 (375 objects). Cluster 2, 1,672 more closely related to formula 2 (206). Cluster 3, 1,591 more closely related to formula 3 (179) and 5 (102). Cluster 4, 1,750 with a greater resemblance to formula 4. The results suggest, therefore, a notable homogeneity in the standard distribution of the fundamental dactyloscopic formulas. The Federal District has a standard percentage that is highly similar to the National figure.

Table 4: Sum of the Residual Square (SQRes)

Phase	Standard	Informed
Location	48.666.16	46.692.89
Relocation	45.143.29	44.286.88

Table 5: Distribution of the objects

Informed K-means (%)							
Cluster	Datiloscopy Formulas						Objetcs
	1	2	3	4	5	others	
1	4.72					37.28	42.00
2		2.59				13.41	16.00
3			2.25		1.28	16.47	20.00
4				1.32		20.68	22.00

Standard K-means (%)							
Cluster	Datiloscopy Formulas						Objetcs
	1	2	3	4	5	others	
1	4.72	2.59	2.25		1.28	16.16	27.00
2				1.32		46.68	48.00
3						16.00	16.00
4						9.00	9.00

5 CONCLUSION

KDD has provided useful results for both researchers and companies alike. In this context, cluster analysis occupies an important place. The clustering process is complex and requires multiple decisions at each stage that influence the final results (Han, 2001). Hanson (1990) highlights the need for a quota of prior knowledge as a requirement for the clustering process. The purpose of the informed clustering is to make use of the information on the data structure and the application domain, not only for purposes of optimizing the operation of the process, but also as restrictions from the data space or the expert's knowledge. This is known as Domain Theory, which guides the inductive process. The structural and causal correlations and dependencies among the attributes can be applied to the homogeneity coefficient. A gradation scale of relevance and interest for the attributes in relation to the desired configuration for the clustering process can also be employed. Research has been driven, furthermore, to apply, the production rules to the homogeneity coefficient of the clustering algorithms. A fertile field for research thus has opened up, especially regarding applications in the area of public safety. The progress of biometric techniques, together with the development of Information Technology, offers interesting possibilities with respect to the problems regarding the civil and criminal identification and differentiation of individuals. According to Araújo (2003), continued research on the database and information of the Federal Police Department's National Identification Institute is significant for boosting the technical criteria used in developing the laws governing civil and criminal identification. Advances in this area of research may also serve to jumpstart and support the development and implementation of computerized fingerprint analysis systems (Automated Fingerprint Identification System – AFIS), thereby generating interesting possibilities for the study of criminal psychology, biology, and anthropology, among other specializations, through improved management and application of the information collected from efforts in the area of civil and criminal identification.

ACKNOWLEDGMENT

We are grateful to the papilloscopist Mr. Marcos Elias Cláudio de Araújo from the INI/DPF who kindly provided the access to the "MECA-Sinic"database.

REFERENCES

Araújo, M. E. C., Bossois L. M., Santana J. L., 2003. *O Arquivo datiloscópico criminal brasileiro: os tipos fundamentais e suas freqüências*. In XIII Congresso Mundial de Criminologia. Sociedade Internacional de Criminologia.

Fayyad, U. M. et al, 1996. From data mining to knowledge discovery: an overview. In: Fayyad, U. M. et al. *Advances in Knowledge discovery and data mining*, AAAI Press. Menlo Park, CA.

Han, J., Kamber, M., 2001. *Data Mining: concepts and techniques*, Morgan Kaufmann Publishers.

Hanson, S. J., 1990. *Conceptual Clustering and Categorization: Bridging The Gap Between Induction and Causal Models.* In: Kodratoff, Y. & Michalski, R.

(Eds.), *Machine Learning: An Artificial Intelligence Approach*, Morgan: San Mateo, CA.

INI - Instituto Nacional de Identificação, 1987. *Identificação Papiloscópica*, Departamento de Polícia Federal (DPF). Brasília.

Oliveira, M. G., 2003. *Otimização de busca decadactilar para métodos manuais, tradicionais ou sistemas automatizados de identificação de impressões digitais (AFIS), utilizando técnicas de Data Mining.* UNB. Brasília.