3rd International Conference on Automation, Control, Engineering and Computer Science (ACECS'16)
Proceedings of Engineering & Technology (PET)
pp. 824-829
Copyright IPCO-2016

The Online Handwriting Recognition Through Mining Frequent Patterns

Chekib Gmati [1], Asma Ksiksi [2], Hamid Amiri [3]

LR-SITI Laboratory, National Engineering School of Tunis, EL Manar University
Le belvedere 1002.Tunis.
[1]chekibgmt2007@gmail.com,[2]asma.ksiksi@gmail.com
[3]hamidlamiri@gmail.com

***Abstract*— This paper presents on-line handwriting recognition method based on the extraction of two types of frequent pattern: frequents closed patterns and frequents maximal patterns. We considered the knowledge Discovery process in Databases (KDD) as a guideline for our approach. We generate signatures representatives of the handwriting identity to be analysed. These signatures are based on the extraction algorithms of frequent patterns and on the transformation of the original database. This allowed us to ensure invariance to scaling, rotation and translation.**

***Keywords*— Knowledge Discovery in Databases (KDD); data mining; frequent patterns; spatial relations; online handwriting recognition**

I. Introduction

In view of the technological developments in communication‘s field and various types of new mobile devices, handwriting recognition is becoming in our days a common purpose and a performance criterion. New application areas are incorporated into tablets and iPhone, note for example handwritten notes and features of gestures in iPhone [4] [5] [8]. In fact, signal’s acquisition is performed by an electronic pen, which comprises micro-sensors, or by a stylus on a sensitive surface, or by coupling pen / screen based on an electromagnetic field as the shelves SAMSUNG Galaxy or IPad and graphics pen Wacom tablets.

The handwriting line (also dynamic) is obtained by a continuous input and is presented in the form of a sequence of points ordered in time scale. Furthermore, temporal representation enable having information about writing speed, its morphology, pen pressure and the order and the direction of the trace. In this case handwriting can be considerate as behaviour over time that eventually describes a form at last lifted the pen. Several approaches have been proposed to solve online handwriting recognition issues from signal pre-processing level to classification.

The property of the signal is taking the form of a path, and designed large databases prompted us to look at some data mining algorithms to search through the paths [12].

According to this reason, we will focus to both types of frequent extraction algorithms that have an interesting property in online handwriting recognition.

II. Online Handwriting

A. *The difficulties of online handwriting recognition*

Variability in handwriting is the major element that reflects the actual difficulty of handwriting recognition. Indeed, variability often presents several forms: the intra-writer (from the same writer), and inter-writer variability (compared to several writers) [7].

In fact, there are other difficulties to be overcome during the signal processing of the handwriting. We cite the ambiguity that we can find between shapes as the lowercase "o", the uppercase letter "O" and the symbol "°" ligatures that connect them and the characters that can represented a source of confusion, or diacritic marks. In addition, time information is very useful for solving recognition difficulties, but we must take into account the direction of writing (the clockwise direction or the reverse).

Also the fast speed writing can cause missing dots in the plots. These missing values affect negatively the recognition of online writing.

B. *Architecture of the online handwriting recognition system*

Online handwriting recognition starts from the acquisition of information that corresponds to tracking the trajectory of the pen tip. The main treatments that characterize the process of handwriting recognition are presented in Fig. 1 [7].

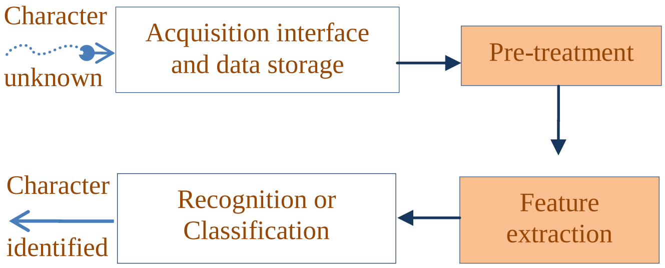
Character
unknown
Acquisition interface and data storage
Pre-treatment
Feature extraction
Recognition or Classification
Character
identified

Fig. 1 General diagram of a system for online handwriting recognition

1) Pre-treatment: The purpose of pre-processing is to eliminate irrelevant information that may adversely affect recognition. Pre-treatment techniques are presented by removing duplicated items, standardization, interpolation of missing points, the peak detection, elimination of unnecessary hooks or small bows and smoothing. In [10] these techniques are applied to the recognition of handwritten Gurmukhi numbers online.

The slow writing can cause the repetition of coordinates at the same position, usually in the corners. All these points are removed by checking the coordinates of any two points are identical. If they are at the same position, one of them is kept and the others are discarded.

The normalization of a trace size has an important role in the online handwriting recognition. A number of nominal size is recognized much more quickly than sales of variable size. Standardization includes basic techniques such as zooming, translation and rotation. Normalization is performed to remove a portion of changes in writing style and simplify online form.

Regarding detection of spikes in a plot, it is based on the calculation of the angles between two successive segments. Through the detection peaks, we can remove unnecessary or small hooks arcs representing a slight deformation of the shape. Finally, smoothing a path is applied in order to eliminate errors from the erratic movement of the hand while writing. This smoothing can be performed by changing the value of each point by the mean value of k neighbours [11].

In [17], the authors highlight the segmentation methods and show the importance of these methods through a state of the art concerning the segmentation in online recognition of Arabic script. They exhibit some segmentation techniques, such as segmentation lines, sub portions letters, or words in the method based on the sign of the slope value. The authors also present the recognition rate of each online system in order to show the contribution of the type of segmentation. In the next section we present a key step in the recognition system, which is linked to the segmentation methods.

2) Features extraction: Types of text recognition characteristics can be classified into three types which are the structural characteristics, the statistical characteristics and the overall transformation. Type shows structural characteristics and describes the geometrical and topological characteristics of a text that can be word, character counts, or path while describing its global and local properties.

We can consider the number of points and their positions, the length of the contour segment, the distance between the start point and end or the projection of the contour on the x-axis and y-axis as a structural characteristic. In general, structural features are difficult to extract from the image of Arabic text and many errors occur because of the similarity of Arabic characters. However, the extraction of these features might be easier and more efficient for the online system because of the way to record the write data. Moreover, this type of feature is commonly used in online recognition systems that have been noticed in the literature. In [18] we find other methods based on the graph, the similarity distance calculation, and many others.

In [12], features are divided in two categories: quantitative and qualitative. The quantitative characteristics of any text quantitative characteristics include the number of points, measures of text, the weight of text above and below the baseline. On the other hand, the quality characteristics include branches, topological descriptions, the topological relationships, the position of the points, connection points, and junction points. The authors present a state of the art approaches made from 1987 to 2011 using several types of features extracted from the signal. Structural characteristics are widely used in relation to statistical characteristics, we can cite as an example the approach based on the Freeman coding [15] which gives a recognition rate of 77% for words. Directions Freeman were used in [16] and gave a rate that rises to 97.6 %.

We note that the approach using the characteristics of global transformation and approaches that combine between structural and statistical characteristics do not meet the recognition rate approaches utilizing channels codes.

We can consider the work on signatures online since writing is a pattern or shape that is built over time. The authors of [17] studied the effect of the use of six overall characteristics of the performance of the verification system. The characteristics are studied: the total time, the total time pen-up, the number of sign changes in the speed of the x and the y direction, the number of sign changes of acceleration in the x-axis and y direction. In online recognition of signatures, the vectors are time-based, the problem occurred when the calculated distance between the similarity vectors which are not the same size.

In [13], the authors used ten parameters: six dynamic characteristics based on the time information and four static characteristics modelling the appearance and the trajectory representing the geometric and spatial information. The author presents in [14] a feature extraction based on an alphanumeric coding according to the directions and the position of the lines.

III. FREQUENT PATTERN MINING

C. Knowledge Discovery in Databases (KDD)

No trivial process of identifying, of valid patterns (for a new data with a high degree of certainty), potentially useful, ultimately understandable. (Should lead to a helpful decisions), ultimately understandable. The overall process consists on transforming the low-level data in high-level data [1]. The process of knowledge discovery in databases (KDD) is composed of several steps presented in Fig. 2.

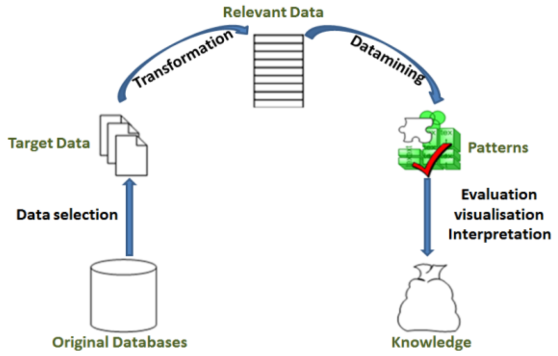

Fig. 2 Knowledge discovery in databases.

The first step called the Data Selection step is composed on a various tasks of integration and cleaning. It consists in identifying and eliminating the noisy data from databases. The second step is a pre-processing step, it select and standardize the relevant data in the decision process, also it allows to reduce the number of dimensions and other well treatments of transformation that can be used as raw data for analysis.

The data mining step occurs to extract patterns or to synthesize the information contained in the data previously treated [2]. Depending on the needs and the objectives of data analysis, models are extracted by different techniques: classification, clustering and or extraction of frequent patterns [1].

D. Frequent patterns

1) Definitions: We consider a set of n items such as I = {i1, i2, …, in}. D is a database with m transactions such as D = {t1, t2,…, tm}, for each ti, with 1≤ i ≤ m, it is composed on an itemset I1 C I (a unique identifier is assigned to each transaction):

- An itemset I1 is a set of attributes or items contained in I. The smallest itemset is the empty set. The largest itemset contains all the items I. An itemset having a size k is called k-itemset.
- The support of an itemset I1 is defined as the percentage of transactions of D containing I1.
- An itemset I1 is said frequent if the support is higher than or equal to the minimum threshold support (minsupp) defined by the user:

$$\text{support(I1)} \geq \text{minsupp} \qquad (1)$$

In the literature several types of frequent patterns were extracted via very specific algorithms, in the following we present only two types: frequent closed patterns and the maximum frequent patterns.

2) Frequent closed patterns: Closed itemsets are a condensed form of all frequent itemsets. Otherwise, we can know the support for each frequent itemset, without scanning the database, based on closed itemsets.

An itemset is said closed if it is frequent and it has no common on-set that has the same support [6] [9].

3) Frequent maximal patterns: An itemset is said maximal if it is frequent and if he has no appropriate frequent superset.

IV. Motivation

Our ultimate goal is to represent the identity of online handwriting without taking account of the writer and with minimal information. More than that, we want a global representation.

Scripters change, variations in writing are notable but still recognizable; in other words the identity of the letters has been preserved. We propose to apply two types of algorithms for extracting frequent patterns: an algorithm for the extraction of frequent closed patterns and an algorithm for the extraction of the frequent maximal patterns.

We chose these two algorithms according to their already mentioned properties and because their compression effects. Indeed, the characteristic of closed frequent patterns is that there is no loss of information in the extracted frequent patterns, which can help us to achieve our goal [3].

Furthermore, the property which characterizes the algorithm of extraction of frequent maximal patterns indicates that a frequent pattern is said maximal if any one of her directly sub pattern is frequent. So we can consider that the frequent patterns are representative or "bounding", which leads us to say that this type of algorithm may be the solution to achieve the identity of handwriting with a less data.

V. Contribution

E. Selection of data

A step of cleaning is crucial to preserve the useful data in the database. This step is to remove all redundant pixels. This redundancy is due to the speed of the writer who does not follow the sampling rate for recording the coordinates of the pixels; then we can find the same pixel recorded several times. This step is essential because we must choose successive pixels in space.

F. Transformation and pre-treatment

In the database we have an ordering list of pixels represented by coordinates X and Y. We will successively extract segments that are represented by four pixels, what are meaning that for each step of four pixels we will extract a segment in chronological order (from time t_0 until the end of recording).

During the extraction of each segment, we will extract the characteristics of each. The characteristics that we must calculate and record are as follows:

- we use an elliptical envelope on each segment to identify the major axis and the minor axis. These two elements described us in an explicit way the curvature and length of each segment. Also, they implicitly communicate to us the relationship between her and the speed of writing (for example, if the minor axis is not around 0 and the major axis is very large then the writer has accelerated in this area trace).
- we will grasp the sense of the order of pixels in the segment. Respecting the order and relying on pixel coordinates, we mentioned that the direction is from bottom to top, top to bottom or the direction is horizontal. This element is very important; it allows us to discriminate the segments with the same characteristics as the other senses. For example, it allows us to neglect the "mirror effect" found in some letters like "N" and "M".
- we will locate the position of the segment that follows current segment. This data will include three values: "right," "left" or "continuity". This data allows us to guarantee the invariance for rotation and scaling invariance. Indeed, the position of the next segment is compared to the current segment, and thus in changing the rotation, data will remain the same. Also during a change of scale, the data is invariant.

When the three features are extracted (with preserving the order) we will discretise the values of the major axes and minor axes on 3 levels. This discretization is achieved by clustering Kmeans algorithm. Then we will get 3 groups for minor axes (small, medium and large) and three groups for the major axes (small, medium and large).

G. Datamining

In this stage the idea is to extract frequent patterns of the new database after transformation (paragraph IV.B). When we have applied the algorithm of frequent pattern on data base, we will obtain a set of frequent patterns related to each word.

Then we will standardize the size of frequent patterns that characterize each word in a matrix. This matrix may be representative by applying to each column the standard deviation. So we get an imprint of the dispersion values. Finally, we have a vector (or signature) representative for each frequent pattern groups (respectively for each word).

We will validate the quality of signatures by a supervised classification with some classification algorithms.

VI. Results

We used a database composed of Latin words. Each word has two different letters; we took into consideration the intra-writer variations and the inter-writer variations.

We used to data entry a graphic tablet MANHATTAN. Data recording is done by the recovery of x and y coordinates and the time t of the entry in an XML file. We applied a supervised classification algorithm using five classifications via Weka: « Multilayer Perceptron », « SVM », « K-nearest neighbour », « J48 » and « Naive Bayes ». Concerning the validation method we used « Cross Validation ».

The used extraction algorithms of frequent patterns are: the Charm algorithm for frequent patterns closed [19] and Charm MFI algorithm for maximal frequent patterns [20].

In Fig. 3 and Fig. 4, the quality of generating signatures after the extraction of frequent closed is clearly higher. The percentage of classification varies between 53.85% and 65.03% for minsupp = 0.65; satisfactory results for J48 and Multilayer Perceptron the results vary between 71.32% and 76.92% for minsupp = 0.4 and = 0.45 minsupp. But the results are higher for minsupp = 0.5, they reach 79% for Multilayer Perceptron and K-nearest neighbor and 74.82% for J48. Values are a little higher than minsupp = 0.45.

We note that the rates are low when the minsupp is very low (0.20%) or very large (0.65%) for the two algorithms. Only the rates were significantly higher for Charm algorithm for minsupp value equal 0.5 to 0.55 (that is compared to the second algorithm Charm MFI).

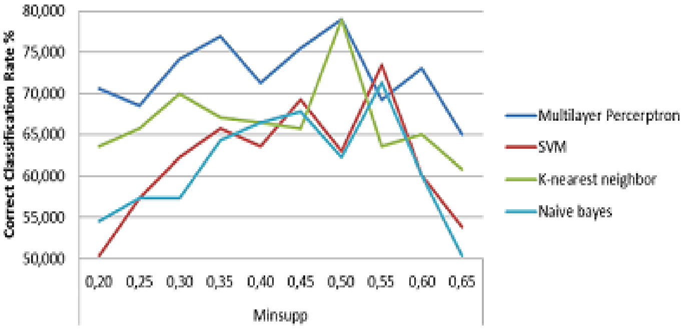

Fig. 3. Variation curves of the correct: Charm algorithm

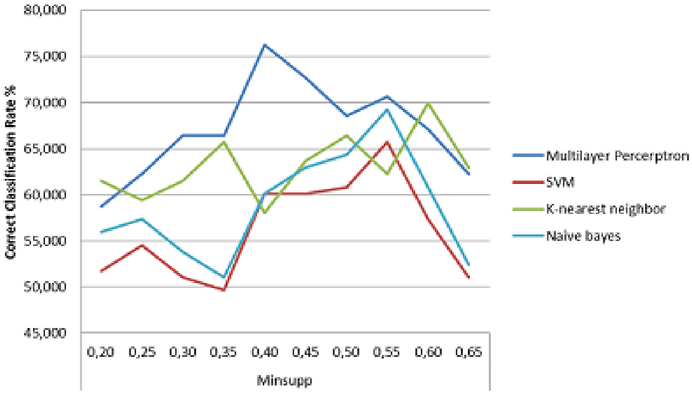

Fig. 4. Variation curves of the correct classification rate: Charm MFI algorithm

We note that the rates are low when the minsupp is very low (0.20%) or very large (0.65%) for the two algorithms. Only the rates were significantly higher for Charm algorithm for minsupp value equal 0.5 to 0.55 (that is compared to the second algorithm Charm MFI).

Although we found a peak value equal to 0.4 minsupp, we cannot say that a signature is representative, this is for two reasons:

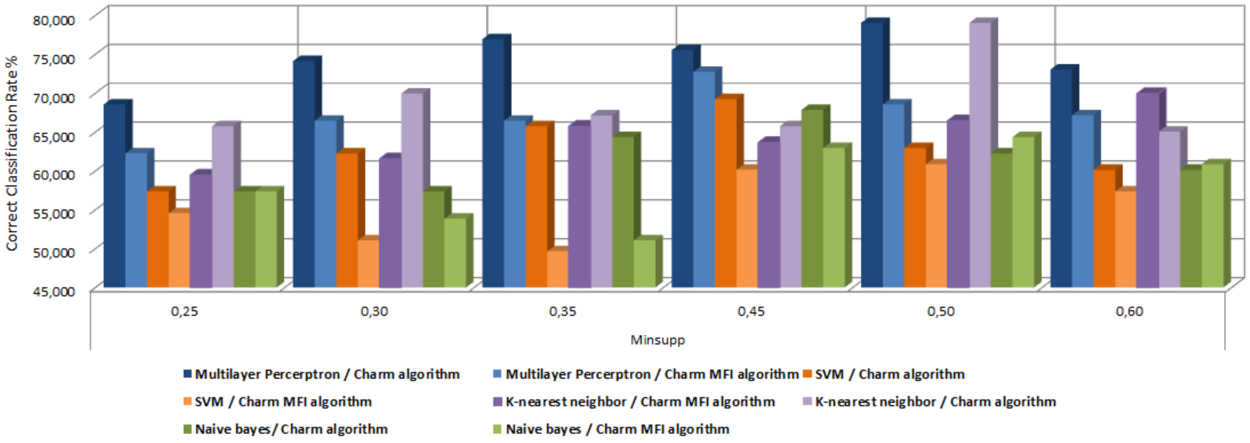

Fig. 5. Variation of the correct classification rate with: Charm MFI and Charm algorithms

- The first is that for the same value 0.4 the correct classification rate is very low for other classification algorithms.
- Secondly, we want the signature represents the identity of words, this is guaranteed at a high value of minsupp, except in this case the value of minsupp is equal to 0.4. So, we determined that this value should be neglected.

The results presented by the charm algorithm, are more convincing (Fig. 5). The loss of information caused by the extraction of the maximum frequent patterns via the Charm MIF algorithm significantly affects the results.

VII. Conclusion

We find that the signatures, based on frequent closed patterns, are best in terms of quality. The signature is extracted from a group of frequent patterns representing the words. The idea is to represent the word in a global way and being insensitive to noise and variations of writing. We find that we must focus on methods that apply the extraction of frequent patterns as association rules taking into consideration the time factor.

References

[1] Sushmitamitra, Tinkuacharya, “DataMining ultimedia, Soft Computing and Bioinformatics”, Wiley-Interscience edition, pp5-6, 2003.

[2] L.Di Jorio, “Recherche de motifs graduels et application aux données médicales”, thesis, University of Montpellier, 2011.

[3] A.Giacometti, D. H. Li, A. Soulet, "20 ans de découverte de motifs : une étude bibliographique quantitative », published in « Revue de nouvelles technologies de l’information », Francophone Conference in Knowledge Extraction and Management (EGC), Toulouse, France,2013.

[4] H.Tu, X.Ren, S.Zhai, “A comparative evaluation of finger and pen stroke gestures”, CHI '12 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2012.

[5] Y.Li, “Gesture search: a tool for fast mobile data access”, Proceeding UIST '10 Proceedings of the 23nd annual ACM symposium on User interface software and technology, pp 87-96, 2010.

[6] B. Négrevergne, J. F. Méhaut, A. Termier, T. Uno, « Découverte d'itemsets fréquents fermés sur architecture multicoeurs », Francophone Conference in Knowledge Extraction and Management (EGC), pp 465-470, Hammamet, Tunisia, 2010.

[7] N.Mezgheni, « Densités de probabilité d’entropie maximale et mémoires associatives pour la reconnaissance en ligne de caractères Arabes », thesis, National Institute of Scientific Research (INRS), Energy, Materials and Telecommunications center, 2005.

[8] J. C. Mathé, « Les entreprises du nouveau monde : 20 histoires – Etudes de cas », Edition of « L’Harmattan », 2010.

[9] K.AOUICHE, « Techniques de fouille de données pour l’optimisation automatique des performances des entrepôts de données », thesis, university of Lumière Lyon, 2005.

[10] R.K.Bawa, R.Rani, « A Preprocessing Technique for Recognition of Online Handwritten Gurmukhi Numerals », International Conference, HPAGC 2011, Chandigarh, India, 2011.

[11] N. Mezghani, A. Mitiche,M. Cheriet, "Bayes Classification of Online Arabic Characters by Gibbs Modeling of Class Conditional Densities", Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol. 30, no. 7, pp. 1121,1131,2008.

[12] M. Z. Khedher, G. A. Abandah and A. M. Al-Khawaldeh, "Optimizing Feature Selection for Recognizing Handwritten Arabic Characters", World Academy of Science, Engineering and Technology, vol. 4, 2005.

[13] N. Tagougui, H. Boubaker,M. Kherallah, A.M. Alimi, "A hybrid MLPNN/HMM recognition system for online Arabic Handwritten script", Computer and Information Technology (WCCIT), vol. 1, no. 6, pp. 22-24, 2013.

[14] Y.Li, J.Wang, “A New Approach to Recognize Online Handwritten NǔShu Characters“, Advances in Intelligent and Soft Computing, vol. 159, pp. 193-199, 2012.

[15] R. I. Elanwar, M. A. Rashwan and S. A. Mashali, "Simultaneous Segmentation and Recognition of Arabic Characters in an Unconstrained on-line Cursive Handwritten Document," International Journal of Computer and Information Science and Engineering, vol. 4, pp. 203-206, 2007.

[16] M. A. H. Omer and M. Shi Long, "Online Arabic Handwriting Character Recognition using Matching Algorithm", published in The 2nd International Conference on Computer and Automation Engineering, pp. 259-262, 2010.

[17] Abuzaraida, M.A.; Zeki, A.M.; Zeki, A.M., "Segmentation techniques for online Arabic handwriting Recognition: A survey," Information and Communication Technology for the Muslim World (ICT4M), pp. 13-14, 2010.

[18] I.M. El-Henawy, M. Z. Rashad, O. Nomir, K. Ahmed, "Online Signature Verification State of the art", International Journal of Computers & Technology , vol. 4, no. 2, 2013.

[19] M. J. Zaki and C.-J. Hsiao. ChARM: An efficient algorithm for closed itemset mining. In 2nd SIAM International Conference on Data Mining, pages 457–473, April 2002.

[20] L. Szathmary, “Symbolic Data Mining Methods with the Coron Platform”, thesis, Henri Poincaré University, Nancy, pp. 58–60, 2002.