A one class KNN for signal identification: a biological case study

Vito Di Gesù, Giosuè Lo Bosco* and Luca Pinello

Dipartimento di Matematica ed Applicazioni, Università di Palermo, Via Archirafi 34, 90123 Palermo, Italy
E-mail: vito.digesu@unipa.it
E-mail: lobosco@unipa.it
E-mail: pinello@unipa.it
*Corresponding author

Abstract: The paper describes an application of a one class KNN to identify different signal patterns embedded in a noise structured background. The problem becomes harder whenever only one pattern is well-represented in the signal; in such cases, one class classifier techniques are more indicated. The classification phase is applied after a preprocessing phase based on a multi layer model (MLM) that provides preliminary signal segmentation in an interval feature space. The one class KNN has been tested on synthetic and real (Saccharomyces cerevisiae) microarray data in the specific problem of DNA nucleosome and linker regions identification. Results have shown, in both cases, a good recognition rate.

Keywords: one class classifiers; multi layer methods; nucleosome positioning.


Biographical notes: Vito Di Gesù is a Full Professor of Computer Science at the Dept. of Mathematics and Applications, University of Palermo. His research interests include data analysis with application to image analysis and bioinformatics. He is the author of a great deal of research studies published at international journals, conference proceedings as well as book chapters.

Giosuè Lo Bosco received his PhD in Mathematics at the University of Palermo in 2004. He is an Assistant Professor at the Dept. of Mathematics and Applications. His research mainly resides in data analysis topics, in particular, clustering and classification of multidimensional data with broad field of applications including bioinformatics.

Luca Pinello received his Masters in Computer Science in 2008. He is a PhD student at the Dept. of Mathematics and Applications. His research is focused on digital signal processing and statistical and computational methodologies for genomic data analysis.
1 Introduction

Classification algorithms are usually based on a training set, where both positive and negative examples are considered. However, in many cases, either only positive examples are available or the two classes are very much unbalanced. This condition may occur in many applications; for example, in Manevitz and Yousef (2001), a one class support vector machine (SVM) is used for the analysis of documents. One class classifiers have been introduced in order to discriminate a target class from the rest of the feature space (Tax, 2001). The approach is based on finding the smallest volume hypersphere (in the feature space) that encloses most of the training data. This approach is mandatory when only examples of the target class are available and it can reach a high performance if the cardinality of the target class is much greater than the other one so that too few training examples of the smallest class are available in order to properly train a classifier. Recently, Wang and Stolfo (2003) have shown that a one class training algorithm can perform equally well as two class training approaches in the specific problem of masquerader intrusion detection. It has also been shown that the combination of several one class classifiers may improve the performance: Nanni (2006) presented an online signature verification system based on one class classifiers, where the recognition is performed by using an ensemble of one class classifiers. Recently, an SVM one class classifier has been designed in Zeng et al. (2006) for the analysis of spontaneous facial expressions in order to discriminate between emotional and non-emotional states. The authors first apply kernel whitening that uses the idea of kernel PCA to map the data into a new space, then, they use support vector data description for the classification to directly fit a boundary with minimal volume around the target data. This is done by finding the centre $C$ and the radius $R$ of a hypersphere that represents the target data; then a new unknown object is classified as a member of the target class if it falls inside the hypersphere.

In this paper, a one class \textit{KNN} (OC-KNN) is proposed in order to identify signal patterns slightly differentiated and embedded in a noise structured background. Examples of such kind of signal are those provided by microarray data of the \textit{Saccharomyces cerevisiae} (Yuan et al., 2005), where the goal, in this case, is the identification of nucleosomes and linker regions across DNA. Nucleosomes are the fundamental repeating units of eukaryotic chromatin and their position can be regulated in vivo by multisubunit chromatin remodelling complexes and can influence gene expression in eukaryotic cells (Corona and Tamkun, 2004; Saha et al., 2006). Alterations in chromatin structure, and hence, in nucleosome organisation, can result in a variety of diseases, including cancer, highlighting the need to achieve a better understanding of the molecular processes modulating chromatin dynamics (Jacobson and Pillus, 1999). To measure the nucleosome positions on a genomic scale, a DNA microarray method has been recently developed. This new approach allows the identification of nucleosomal and linker DNA sequences on the basis of susceptibility of linker DNA to micrococcal nuclease (Yuan et al., 2005). This method allows the representation of microarray data as a signal of green/red ratio values that shows nucleosomes as peaks of about 150 base pairs long, surrounded by lower ratio values corresponding to linker regions. Note that, because of the noise in the data signal and the large-scale trends in mean hybridisation values, a naive threshold-based approach for determining nucleosome positions is considered highly inaccurate. Therefore, a hidden Markov model (HMM) approach has been used to
discriminate regions representing nucleosomes and linkers. However, this approach suffers from constrain imposed by its static topology, as a consequence lots of potential good input data are discarded from the analysis. Moreover, this approach does not take into account the shape information of the green/red ratio values.

In particular, this specific biological problem concerns the discrimination of two classes of patterns: the well-positioned nucleosomes (WPN) and the linkers (LN). For an example of an input signal which also shows the two pattern classes, see Figure 1(a).

**Figure 1** (a) Input signal* (b) pattern identification and extraction**

Notes: *The input signal is the logarithmic ratio of the green channel to red channel values for each spot of the microarray. Nucleosomes correspond to peaks (marked by black circle), surrounded by lower ratio values corresponding to linker regions (marked by dashed circles) that are nucleosome free.

**In this example, six patterns are retrieved, identified by rhombus, circle, square, triangle down, triangle up and a star. Each pattern identifier is replicated for each of its feature values and pointed in each one of its middle point.
The OC-KNN is applied after a preprocessing phase, based on a multi layer model (MLM) (Corona et al., 2007) that provides preliminary signal segmentation in an interval feature space. In this work the OC-KNN has been tested on real and synthetic microarray data. Results have shown a good recognition rate on real and simulated data for nucleosome and linker regions. The paper is organised as follow: Section 2 provides a review of related works; Section 3 describes the OC-KNN classifier and its assessment; the MLM preprocessing phase is outlined in Section 4; in Section 5 the adaptation of the MLM on a particular biological problem is outlined and the results of the OC-KNN on synthetic and real biological data are shown; final remarks and discussion are given in Section 6.

2 Related papers

The first algorithms for one class classification were based on neural networks, such as those of Moya et al. (1993), Moya and Hush (1996) and Japkowicz et al. (1995).

More recently, one class versions of the SVM have been proposed by Scholkopf et al. (2001). The aim is to find a binary function that takes the value +1 in a small region capturing most of the data, and –1 elsewhere. Data transformations are applied such that the origin represents outliers, then the maximum margin, separating hyperplane between the data and the origin, is searched.

The application of machine learning to classification problems, which depends only on the positive examples, is gaining attention in the computational biology community. In this section, we list some applications of one class classifiers to biological and biomedical data.

In Yousef et al. (2008), a study using one class machine learning for microRNA (miRNA) discovery is presented. Authors compare an OC-KNN to two class approaches using naive Bayes and SVMs. Using the EBV genome as an external validation of the method, they found one class machine learning to work as well as or better than a two class approach in identifying true miRNAs as well as predicting new miRNAs.

In Nanni (2005), a general method for predicting protein-protein interactions is presented. The search of feasible interactions is carried out by a learning system based on experimentally validated protein-protein interactions in the human gastric bacterium Helicobacter pylori. The author shows that the linear combination of discriminant classifier provides a low error rate.

In Pekalska et al. (2004), a one class classification problem is applied to the detection of diseased mucosa in oral cavity. Authors either combine several measures of dissimilarity of an element from a set of target examples in a single one class classifier or combine several one class classifiers trained with a given measure of dissimilarity. Results show that both approaches achieve a significant improvement in performance.

3 One class KNN

A KNN classifier for $M$ classes problem is based on a training set $T$ for each class $m$, $1 \leq m \leq M$. The assignment rule for an unclassified element $x \in X$ is:

$$j = \arg \max_{1 \leq m \leq M} K^{(m)}(x)$$

(1)
where \( \sum_{K}^{(m)} (x) \) are the training elements of class \( m \) in the \( K \) nearest neighbours of \( x \). One of the crucial points of the \( KNN \) is the choice of the best \( K \), which is usually obtained minimising the misclassification rate in validation data.

In the case of a binary classification \( (M = 2) \), one class training means that in the decision rule, we use the training examples of only one class. Here, we propose a one class training \( KNN \) (\( OC-KNN \)) which is a generalisation of the classical \( KNN \) classifier (Jain and Dubes, 1988). Let \( T_p \) be the training set for a generic pattern \( p \) representing a positive instance and \( \delta \) a dissimilarity function between patterns. Then the membership for an unknown pattern \( x \) is:

\[
\chi_{\phi,K}(x) = \begin{cases} 
1 & \text{if } \left| \{ y \in T_p \text{ such that } \delta(y,x) \leq \phi \} \right| \geq K \\
0 & \text{otherwise} 
\end{cases}
\]  

(2)

Informally, the rule says that if there are at least \( K \) patterns in \( T_p \) dissimilar from \( x \) at most \( \phi \), then \( x \) is supposed to be a positive pattern, otherwise it is negative. It can be simply proved that the \( OC-KNN \) has some interesting properties:

**Proposition 3.1:** Let \( D \) a dataset of patterns, \( T_p \subseteq D \) the training set for the positives, \( S_{\phi,K} = \{ x \in D | \chi_{\phi,K}(x) = 1 \} \) the set with membership \( \chi_{\phi,K} \), then:

a) \( S_{\phi,K'} \subseteq S_{\phi,K} \forall K' \geq K \)

b) \( S_{\phi,K} \subseteq S_{\phi',K} \forall \phi' \leq \phi \)

The \( OC-KNN \) performance depends on the threshold, \( \phi \), and the number of neighbours, \( K \), that are used in the classification phase. Both of them can be determined by using a validation procedure applied on the training set of positives \( T_p \). In the following, we describe the procedure used to estimate the best pair \( (\phi^*,K^*) \).

Let us define the performance function \( M \):

\[
M(\phi,K) = \frac{|S_{\phi,K}|}{|T_p|}
\]

Note that, in this validation procedure \( \forall x \in T_p \) assigned to \( S_{\phi,K} \), use the membership \( \chi_{\phi,K}(x) \) defined on the training set \( T_p = \{ x \} \). By using \( M \), we can define the functions \( P \) and \( Q \):

\[
P(\phi) = \sum_{k \in \{K_n,K_m\}} M(\phi,k) \quad \text{and} \quad Q(k) = \sum_{\phi \in \{\phi_m,\phi_M\}} M(\phi,k)
\]

where \( \{\phi_m,\phi_M\} \) and \( \{K_n,K_M\} \) are sets of increasing values of thresholds and number of neighbours, respectively. By applying Proposition 3.1, it results that the function \( M \) increases while the threshold \( \phi \) increases, and decreases while the neighbours \( K \) increases. In Figure 2(a), a 3D plot of the function \( M \) relative to the classification of nucleosome and linker regions on the \( Saccharomyces cerevisiae \) dataset is shown.
Assigning the values, $\phi_m = \min_{x,y \in T_p} \delta(x,y)$ and $\phi_M = \max_{x,y \in T_p} \delta(x,y)$, $K_m = 1$, $K_M = |T_p|$, the pair $(\phi^*, K^*)$ to choose is:

$$\phi^* = \min \{ \phi | P(\phi) = \max \{ P(\phi) \} \}$$
$$K^* = \max \{ K | Q(K) \neq 0 \}$$

Informally, such estimation methodology selects the smallest threshold $\phi^*$ which causes the best performances on the validation data, most independently from the values of $K$. Moreover, the value $K^*$ is chosen to be the largest one causing performances different from zero. In Figure 2(b), an image representation of $M$ shows also the chosen $(\phi^*, K^*)$ concerning the classification of nucleosome and linker regions on the *Saccharomyces cerevisiae* dataset. A fuzzy extension version of the OC-KNN has been recently tested on two public datasets (Di Gesù and Lo Bosco, 2007), studying also the gain in classification performances when combining several one class classifiers defined by different dissimilarity functions.

**Figure 2** Two different representation of $M$. (a) a 3D plot (b) an image representation showing the values of $M$ using greyscale (0 is black, 1 is white)

Notes: In this latter figure, there is also the chosen pair $(\phi^*, K^*)$.

### 4 Multi layer model

In this section, an outline of the MLM for the analysis of monodimensional signal is provided. It can be considered a preliminary step that provides to the OC-KNN the proper feature space and input pattern to be classified. The MLM procedure is carried out as follows:
Preprocessing: A preprocessing is necessary in order to reduce the effect of the signal noise. Starting from the input signal, \( S \), each fragment \( S_t, 1 \leq t \leq T \), is convolved by a generic kernel window. After this process, \( X \) represents the convolved signal.

Training set and model construction: A model set \( M = \{ X_t \mid X_t \subseteq X \} \) is built by extracting subsignals \( X_t \) of the convolved signal \( X \) that satisfy a particular set of conditions. Such conditions are defined with respect to the shape of the pattern we want to identify. For example, in the case of the signals resulting from tiling microarray data, we select a particular bell-shaped subsignals centred on the local maxima of the input signal.

Interval identification: The core of the method is the interval identification by considering \( H \) threshold levels \( t_h (h = 1, \ldots, H) \) of the convolved signal \( X \). For each \( t_h \) a set of intervals \( R_h = \{ I_h^1, I_h^2, \ldots, I_h^H \} \) is obtained; where \( I_h^j = [b_h^j, c_h^j] \), where \( b_h^j, c_h^j \) are the lower and upper limits of the interval and \( X(b_h^j) = X(c_h^j) = t_h \).

Interval merging, pattern definition and selection: This step is performed by taking into account again the shape of the pattern to classify. In particular, a set of rules \( R \) on the intervals extracted in the previous step is defined respecting several conditions inspired from the knowledge of the pattern to search (shape, persistence, etc.). The application of such rules on the set of intervals \( R = \{ R_h \mid 1 \leq h \leq H \} \) groups such intervals, defining the patterns:

\[
P_i = \left\{ I_{j_1}^1, I_{j_1}^2, \ldots, I_{j_1}^l \right\} \mid R \left( I_{j_1}^1, I_{j_1}^2, \ldots, I_{j_1}^l \right) \text{is satisfied} \right\}
\]

Feature extraction: Each pattern \( P_i \) is identified by \( I_{j_1}^1, \ldots, I_{j_l}^l \). Straightforwardly, the feature vector of \( P_i \) is a \( 2 \times l \) matrix where each column represents the lower and upper limits of each interval from the lower threshold \( j \) to the upper threshold \( j + l \). The representation in this multidimensional feature space is used to characterise different types of patterns.

Dissimilarity function: A dissimilarity function between patterns is defined in order to measure their dissimilarity. This is fundamental in the case of classification.

5 Experiments and results

The OC-KNN has been tested on monodimensional signals provided by microarray data, where the goal, in this case, is the identification of nucleosomes and linker regions across DNA. The used microarray method allows the representation of data as a signal of green/red ratio values that shows nucleosomes as peaks of about 150 base pairs long, surrounded by lower ratio values corresponding to linker regions (Yuan et al., 2005).
5.1 MLM rules on biological data

The key steps of the MLM are the training set and model construction and the interval merging, pattern definition and selection, which have to be defined in a personalised way depending on the property of the pattern to retrieve. In the following, these steps are defined in the case of the monodimensional signal showing nucleosome positioning information. Note that, in such case, the preprocessing phase consists in a convolution by an averaging kernel window $w = \left[\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right]$.

5.1.1 Training set and model construction

Since we know that WPN are shown as peaks of a bell-shaped curve, in order to locate the position of a nucleosome, all local maxima of the input signal are automatically extracted from the convolved signal $X$ of $S$. Each convolved fragment $X_i$ is processed in order to find $L(X_i)$ local maxima $M_i^{(l)}$ for $l = 1, \cdots, L(X_i)$. The extraction of each subfragment for each $M_i^{(l)}$ is performed by assigning all values in a window of radius $\text{os}$ centred in $M_i^{(l)}$ to a vector $F_i^l$ of size $2 \times \text{os} + 1$: $F_i^l(j) = X_i(M_i^{(l)} - \text{os} + j - 1)$, for $j = 1, 2, \ldots, 2 \times \text{os} + 1$. Note that the $\text{os}$ value is related to the extent of the bell curve representing a nucleosome. The selection of the significant subfragments – to be used in the model definition – is performed by satisfying the following rule:

$$
\begin{cases}
F_i^l(j + 1) - F_i^l(j) > 0 & j = 1, \cdots, \text{os} \\
F_i^l(j + 1) - F_i^l(j) < 0 & j = \text{os} + 1, \cdots, 2 \times \text{os}
\end{cases}
$$

Note that such rule is always satisfied for signal fragments representing bell-shaped curves. After the selection process $G(X_i)$ subfragments remain for each $X_i$. The model of the interesting pattern is then defined by considering the following average:

$$
F(j) = \frac{1}{T} \sum_{i=1}^{T} \sum_{k=1}^{G(X_i)} F_i^k(j) \quad j = 1, \cdots, 2 \times \text{os} + 1
$$

that is, for each $j$, the average value of all the subfragments satisfying equation (3). The training set of the interesting pattern is $T_p = \{F(j) \mid 1 \leq t \leq T, 1 \leq l \leq G(X_i)\}$.

5.1.2 Interval merging and pattern definition

This step is performed by taking in account that bell-shaped pattern must be extracted for the classification phase. Such kind of patterns are characterised by sequences of intervals $\{I_j^1, \cdots, I_j^n\}$ such that $I_j^1 \supseteq I_j^{n+1}$; more formally a pattern $P_t$ is defined as:

$$
P_t = \{I_j^1, I_j^{n+1}, \cdots, I_j^m \mid j + l \leq H \wedge \forall I_h^k \exists! I \in R_{h+1} : I = I_h^{k+1} \supseteq I_h^k\}
$$
where \( j \) defines the threshold, \( t_j \), of the widest interval of the pattern. From the previous definition, it follows that \( P_i \) is built by adding an interval \( I_{ji+1}^{h} \) only if it is the unique in \( R_{h+1} \) that includes \( I_{ji}^{h} \). Note that, this criterion is inspired by the consideration that a nucleosome is identified by bell-shaped fragment of the signal (thus satisfying equation (3)), and the intersection of such fragment with horizontal threshold lines results on a sequence of nested intervals.

5.1.3 Pattern selection

In this step, the interesting patterns \( P^{(m)} \) are selected following the criterion:

\[
P^{(m)} = \{ P : |P| > m \}
\]

i.e., patterns containing intervals that persist at least for \( m \) increasing thresholds. This further selection criterion is related to the height of the bell-shaped fragment, in fact a small value of \( m \) could represent noise rather than nucleosomes. The value \( m \) is said the minimum number of permanence. In Corona et al. (2007), a calibration procedure able to estimate the value of \( m \) on the same biological problem is described.

5.1.4 Dissimilarity function

A dissimilarity function between patterns is defined in order to characterise their shape:

\[
\delta(P_r, P_s) = \sum_{h=1}^{H} \left| a_h^r - a_h^s \right|
\]

where \( a_h^r = c_h^r - b_h^r \), \( a_h^s = c_h^s - b_h^s \).

Note that a generic pattern \( P \) could not be defined for all the \( h = 1, \ldots, H \). The dissimilarity \( \delta \) is then calculated after completing the feature vector of patterns \( P_r \) and \( P_s \) to \( H \) components by using the pattern model \( F \) in the case a component is not defined. In particular, this dissimilarity is a distance that takes into account the shape of a pattern. An example of the signal and the relative interesting patterns is given in Figure 1(b).

5.2 Synthetic generation of biological signals

A procedure to generate synthetic signal, resulting from the evolution of a first prototype introduced in Corona et al. (2007), has been developed allowing us to assess the feasibility of our method on controlled data. Comparing to the previous version, by using more parameters, it allows a more reliable generation of synthetic microarray data. Generated signals emulate the one coming from a tiling microarray where each spot represents a probe \( i \) of resolution \( r \) base pairs overlapping \( o \) base pairs with probe \( i+1 \). In particular, the chromosome is spanned by moving a window (probe) \( i \) of width \( r \) base pairs from left to right, measuring both the percentage of mononucleosomal DNA \( G_i \) (green channel) and whole genomic DNA \( R_i \) (red channel) within such window, respecting also that two consecutive windows (probes) have an overlap of \( o \)
A one class KNN for signal identification

base pairs. The resulting signal $V(i)$ for each probe $i$ is the logarithmic ratio of the green channel $G_i$ to red channel $R_i$. Intuitively, nucleosomes presence is related to peaks of $V$ which correspond to higher logarithmic ratio values, while lower ratio values shows nucleosome-free regions called linker regions. This genomic tiling microarray approach takes inspiration from the work of Yuan et al. (2005) where the authors have used the same methodology on the *Saccharomyces cerevisiae* DNA. Here we have defined a model able to generate such signals characterised by the following parameters:

- $nn$: The number of nucleosomes we want to add to the synthetic signal.
- $nl$: The length of a nucleosome (we know that in real case a nucleosome is 150 base pairs long).
- $\lambda$: Mean of the Poisson distribution used to model the expected distances between adjacent nucleosomes.
- $r$: The resolution of a single microarray probe.
- $o$: The length in base pairs of the overlapping zone between two consecutive probes.
- $nr$: The number of spotted copies (replicates) of nucleosomal and genomic DNA on each probe of the microarray.
- $dp$: The percentage of the delocalised nucleosomes over the total number of nucleosomes.
- $dr$: The range which limits the delocalisation of a nucleosome in each copy of $nr$. It is defined in base pairs.
- $nsv$: The variance of the green channel in each probe, even in absence of nucleosomes due to the cross hybridisation. This variance follows a normal distribution with mean 0.1.
- $pur$: The percentage of DNA purification, which is the probability that each single DNA fragment of the $nr$ copies appears in the microarray hybridisation.
- $ra$: Relative abundance between nucleosomal and genomic DNA.
- $SNR$: The linear signal to noise ratio of the synthetic signal to generate. Note that the noise is assumed to be Gaussian.

Initially, a binary mask signal $M$ is generated by considering 1 as all the base pairs representing a nucleosome (the nucleosomal regions) and 0 as the regions representing linkers (the linker regions). Note that, the beginning of each nucleosomal region is established by the Poisson distribution with mean $\lambda$. The mask signal $M$ will be used in order to validate the classification results.

The red channel of the microarray (the genomic channel) results from the generation of $nr$ replicates $I^R_1, \ldots, I^R_{nr}$ each one starting from an initial nucleosomal region of random size $b \sim U(0, r)$ (uniformly distributed in the range $[0, r]$), followed by continuous nucleosomal region of $r$ base pairs.

Conversely, in order to simulate the green channel (the nucleosomic channel) $nr$ replicates, $I^G_1, \ldots, I^G_{nr}$ are considered, each one initially equal to $M$ and subsequently
modified by perturbing each starting points $x_D^i$ of the nucleosome to consider as delocalised such that $x_D^i = x_D^i + \mu$ with random $\mu \sim U(dr)$. Note that the percentage of nucleosomes to consider as delocalised is established by the parameter $dp$.

Afterwards, each nucleosomal region on the generic replicate $I_i^R$ and $I_i^G$ can be switched off depending on the value of a random variable $\alpha \sim U(0,1)$. Precisely, each nucleosomal region verifying the test $\alpha < pur$ is considered and set to 1, otherwise it is not considered and set to 0. This results in new replicates $T_i^{R_T}$ and $T_i^{G_T}$. Finally, the generated synthetic signal $V$ for a probe $i$ is so defined:

$$V(i) = \left\{ \log_2 \left[ \sum_{j=1}^{nR} \frac{T_j^{G_T}(k) \cdot ra}{T_j^{R_T}(k)} + \epsilon \right] \right\}_{(r-o)i-r+o+1 \leq k \leq (r-o)i+o}$$

(5)

where $\epsilon \sim N(0,1,nsv)$.

5.3 Results

The following experiments have been carried out by measuring the correspondence between nucleosome and linker regions. In the case of the synthetic signal, the output of the classifier has been compared with a mask $M'$ derived from $M$; in the case of the real dataset, it has been compared with the output of the HMM used in the paper of Yuan et al. (2005) optimally converted into a binary string.

Moreover, by biological consideration, the radius $os$ has been set to $os = 4$. The performances have been evaluated in terms of recognition accuracy, RA. The RA uses a new mask $M'$ obtained by converting $M$ into probe coordinates such that a probe value is set to 1 (e.g., shows a nucleosome portion) if the corresponding base pairs in $M$ include at least 1. The real nucleosomal (linker) regions $RNR$ ($RLR$) are represented by $M'$ as contiguous sequence of 1’s or 0’s, respectively; here we consider that a nucleosomal (linker) region $CNR$ ($CLR$) has been classified correctly if there is a match of at least $0.7 \times L$ contiguous 1’s (0’s) between $CNR$ ($CLR$) and the corresponding $RNR$ ($RLR$) in $M'$ where $L$ is the length $RNR$ ($RLR$). The value 0.7 has been chosen because it represents a 70% of regions overlap very unlikely to be due to chance.

5.3.1 Results on synthetic data

Such experiments allow testing the robustness of the OC-KNN to signal noise. All parameters used in the generation of synthetic data have been inspired by biological considerations and are $nn = 200, nl = 250, \lambda = 200, r = 50, o = 20, nr = 100, dp = 0, dr = 0, pur = 0.8, nsv = 0.01, SNR = \{1,2,4,6,8,10\}$ and $ra = 4$, resulting in six synthetic signals at different $SNR$. The training set $T_p$ is represented by all $WPN$'s that fit better the conditions in equation (3) with $os = 4$, because, by biological consideration, we know
that a nucleosome is around 150 base pairs which corresponds to eight probes. Thus, the training set \( T_p \) and consequently its size \( TL \), are automatically selected by the MLM depending on the generated input signal, resulting that, for the specific experiments reported here, \( TL = \{63,98,127,142,145,147\} \) for \( SNR = \{1,2,4,6,8,10\} \), respectively. The optimal parameters for the MLM are derived by a calibration phase described in Corona et al. (2007) and have resulted \( H = 20 \) and \( m = 5 \). The performances have been evaluated measuring the correspondence between the classified WPN or LN regions and the ones imposed in the generated signal. The parameters \( (\phi^*, K^*) \) of the OC-KNN has been chosen by the validation procedure described in Section 3 for each \( SNR = \{1,2,4,6,8,10\} \). Figure 3 reports the best accuracy and FPR values versus SNR, showing also, for each SNR signal, the \( (\phi^*, K^*) \) causing such values. From this study, it results that the average accuracy and FPR over the six experiments is 94% and 9%, respectively.

5.3.2 Results on real data

In this experiment, we have compared the accordance of the HMM used in the paper of Yuan et al. (2005) on the *Saccharomyces cerevisiae* real data. The input signal representing this data is composed by 215 contiguous fragments for a total of 24,167 base pairs. The training set \( T_p \) has been decided in the same way as above. In such experiment, we have chosen \( H = 40 \), \( m = 6 \) by a calibration phase \( (m = 0.15 \times 40) \) that is fully described in Corona et al. (2007). The confusion matrices, which show the RA of HMM considering MLM as the truth classification and RA of MLM considering HMM
as the truth classification, are reported in Table 1. The results can be summarised in an overall RA of (0.76) for the HMM (MLM true) and 0.65 for MLM (HMM true). In particular, from these studies we can conclude that MLM does not fully agree with HMM on the nucleosome patterns, which indicates that the integration of the two methods could improve the overall classification.

<table>
<thead>
<tr>
<th>H</th>
<th>L</th>
<th>M</th>
<th>H</th>
<th>M</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>L</td>
<td>0.66</td>
<td>L</td>
<td>L</td>
<td>0.65</td>
</tr>
<tr>
<td>M</td>
<td>N</td>
<td>0.14</td>
<td>M</td>
<td>N</td>
<td>0.34</td>
</tr>
</tbody>
</table>

Note: The table on the left shows the RA results of HMM when considering MLM as the truth classification, while the opposite is shown on the right table.

6 Final remarks

In this paper, we have shown that the OC-KNN classifier, by using the preprocessing of the MLM, is able to distinguish between nucleosome and linker patterns in the particular problem of the nucleosome positioning. We have validated our results on simulated dataset, using a method to generate synthetic microarray experiments and on real dataset too. In particular, the results have shown an average accuracy of 94% on six simulated signals generated at different signal to noise ratios and an accordance of 65% with a HMM method devoted to the same purpose. Future work will be devoted to apply of the whole method (MLM and one class classifier) to complex structured data as, for example, on microarray data of the drosophila chromosome and also on electrocardiogram signals.

References


