Applied Computing and Informatics (2014) **10**, 52–67

King Saud University

Applied Computing and Informatics

www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE (INVITED)

On classification in the case of a medical data set with a complicated distribution

Martti Juhola [a,*], Henry Joutsijoki [a], Heikki Aalto [b], Timo P. Hirvonen [b]

[a] *Computer Science, School of Information Sciences, University of Tampere, Tampere 33014, Finland*
[b] *Department of Otorhinolaryngology & Head and Neck Surgery, University of Helsinki and Helsinki University Central Hospital, Helsinki 00029 HUS, Finland*

Received 27 February 2014; accepted 14 March 2014
Available online 25 March 2014

KEYWORDS

Data mining;
Data cleaning;
Complicated data distributions;
Classification

Abstract In one of our earlier studies we noticed how straightforward cleaning of our medical data set impaired its classification results considerably with some machine learning methods, but not all of them, unexpectedly and against intuition compared to the original situation without any data cleaning. After a more precise exploration of the data, we found that the reason was the complicated variable distribution of the data although there were only two classes in it. In addition to a straightforward data cleaning method, we used an efficient way called neighbourhood cleaning that solved the problem and improved our classification accuracies 5–10%, at their best, up to 95% of all test cases. This shows how important it is first very carefully to study distributions of data sets to be classified and use different cleaning techniques in order to obtain best classification results.

* Corresponding author. Tel.: +358 40 1901716; fax: +358 3 2191001.
E-mail address: Martti.Juhola@sis.uta.fi (M. Juhola).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

2210-8327
http://dx.doi.org/10.1016/j.aci.2014.03.001

1. Introduction

In our earlier research we developed a signal analysis method for nystagmic eye movements investigated in otoneurological tests (Juhola et al., 2009, 2011). For the automatic analysis of such signals poor or invalid nystagmic eye movements should correctly be separated from valid nystagmic eye movements, because valid eye movements can only be used for the data analysis needed for the diagnostics of otoneurological patients. Typically, invalid nystagmic eye movements are corrupted by noise or artefacts. Thus, we have also studied the classification of nystagmic eye movement candidates into invalid and valid, hereafter called the rejected and accepted, on the basis of machine learning methods (Juhola et al., 2013). We then observed how their complicated distribution made the classification task difficult and attempted to reduce the greater subset (class) of the rejected eye movement candidates in a learning set, which was performed by cleaning away a part from them. Surprisingly, a simple cleaning process impaired classification results of some of the machine learning methods applied. We realized that the reason for such a seemingly conflicting situation originated from the complicated variable distribution of the data (Juhola et al., 2013).

In order to define which distribution of two classes is seen as simple or complicated we refer to Fig. 1. A simple distribution is where the centre (computed as means for all variables) of each class is located inside its own area including most elements of that class. This is described in Fig. 1(a). A complicated distribution is depicted in Fig. 1(b), where the centre of one class is outside its own area. Such complicatedness could be defined in various ways, but it is essential that we cannot then base data cleaning on distances from the class centres.

Nystagmus is formed from repeated, reflexive eye movements occurring as to-and-for beats that can be measured in the horizontal, vertical and torsional directions with two eye movement video cameras, one for each eye. A hypothetic nystagmic eye movement beat is seen in Fig. 2 and an actual signal of several nystagmic beats in Fig. 3. A nystagmic beat includes the slow phase immediately followed by the fast phase in order to return the eye in the opposite direction. Nystagmic beats are repetitive and their configurations vary even in the course of short measurement times. A healthy subject performs nystagmic eye movements, for example, when he or she is sitting in a moving train and looks at changing (relatively close) views through a window. This is called optokinetic nystagmus because of the stimulation. Caloric nystagmus is induced by injecting a small quantity of cool or warm (37 ± 7 °C) air or water into the ear canal of a subject. In regard to some otoneurological disorders or diseases, head shaking or head movement can provoke nystagmus and even spontaneous nystagmus may appear in vestibular patients. Congenital nystagmus also exists (Hertle and Dell'Osso, 1999). The slow phase features (variables) of nystagmus are important for the diagnostics of vestibular neuritis, positional vertigo, vestibular schwannoma and Menière's disease. The fast phase features of nystagmus are important for

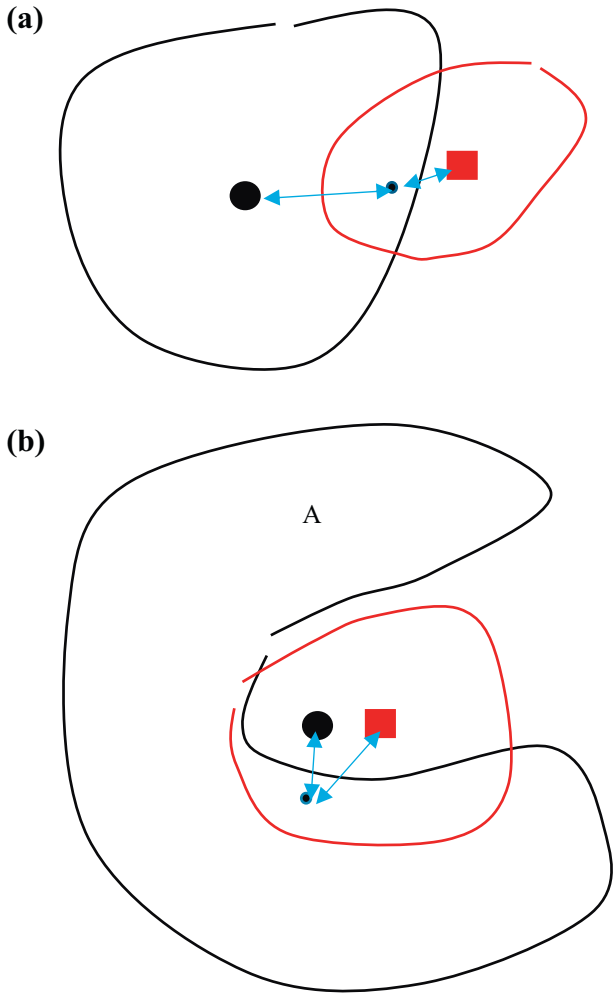

Fig. 1 Let us assume that there are two hypothetical classes A and B in a variable space. Elements of the majority class A close to B and possibly from their overlapping area would be useful to be cleaned out. (a) A simple distribution in which the majority class A (more elements) and minority class B partially overlap and in which element x from A is closer to the centre of B (square) than to the centre (circle) of its own class A. Thus, x can clearly be cleaned on the basis of the distance criterion. (b) A complicated distribution in which x from A is closer to its own centre than to that of B. Further, the centre of A is outside its actual area. Element x cannot be cleaned on the basis of the distance criterion.

investigations of the central origin, i.e., the brain. Our current data included measurements from 107 patients mainly with spontaneous nystagmus. Here the slow phase variables were medically essential.

Since the 1970s, dozens of different nystagmus detection algorithms have been presented (Abel et al., 2008; Augustyniak, 1996; Hertle and Dell'Osso, 1999; Hosokawa et al., 2004; Juhola, 1988; Tominaga and Tanaka, 2010; Wall and Black, 1982). Most of them have been on the basis of applying digital filters,

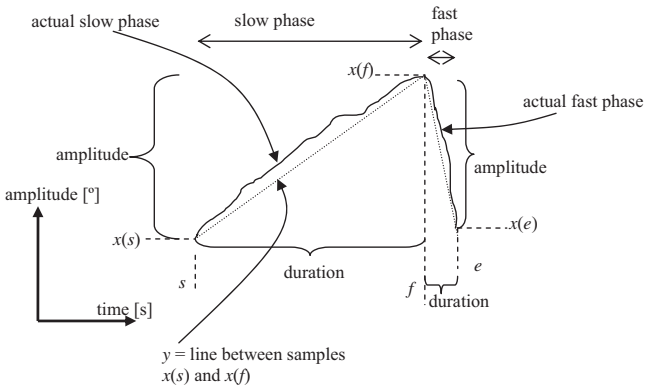

Fig. 2 A hypothetic nystagmic beat describing its main variables and shape from which some other variables are derived. A beat is composed of the slow and fast phases represented with their amplitudes given in angular degrees and durations in seconds. In reality, there may be noise or other unevenness in nystagmus signals. Such phenomena are estimated by computing, for example, linear correlation between an actual slow phase and its ideal component, straight line *y*.

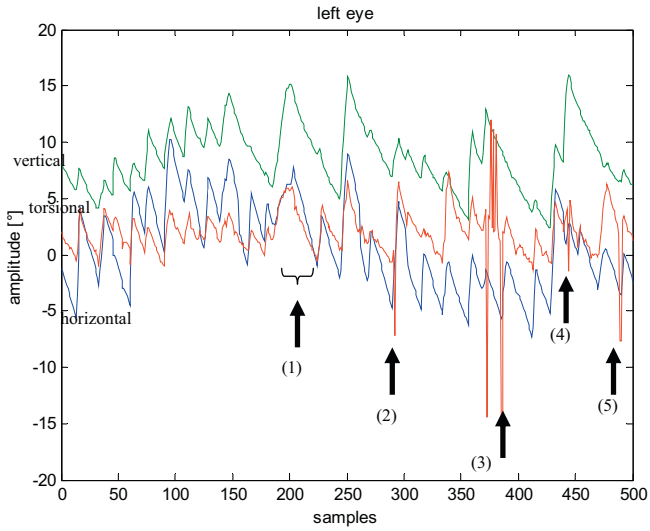

Fig. 3 Here a three-dimensional nystagmic eye movement signal is depicted as three one-dimensional component signals so that their features can be visualized and understood. The component signals are horizontal (blue), vertical (green) and torsional (red). The signal is 10 s long sampled at 50 Hz. Segments (1)-(5) were deemed due to be rejected nystagmic beats because of signal corruption. Segment (1) was corrupted both in the horizontal and torsional components. Segments from (2) to (5) were dropouts of video camera images in the torsional component, in other words, the camera system had momentarily failed to identify the eye in its successive images. Peaks with lower amplitudes than 1° were seen noise or unevenness being neither valid nor invalid nystagmic beats.

thresholding, other signal analysis and tuning parameters. Nevertheless, it is not only to detect nystagmic beat candidates, but the separation of acceptable beats from those corrupted or noisy is a necessary stage in processing before computing nystagmus variable values from all accepted beats. Typically, this stage has been grounded on conventional signal analysis methods and separately on every single nystagmic eye movement. However, this is difficult, because poor nystagmic beats may vary remarkably between patients and also depend on measurement devices.

Recently, we developed an approach based on machine learning and collected a data set of nystagmic eye movement beats, both accepted and rejected, in order to form a training set for machine learning and classification of nystagmus signals (Juhola et al., 2013). To our knowledge, no such attempt has previously been made for nystagmus data. Nonetheless, the classification task was difficult because the data distribution appeared to be of a rare form in which the two classes of the data were very close to each other or partially overlapped so that accepted nystagmic beats were, in a way, surrounded by those rejected beat candidates (Juhola et al., 2013). This resulted in an extraordinary outcome that our then straightforward data cleaning technique impaired the classification results computed with some machine learning methods. In the present study we applied a more efficient method (Laurikkala, 2001) to overcome this problem.

In the present description we do not express in detail how nystagmic beat candidates were detected and how they were determined to be acceptable, because all these are preprocessing phases for the study to be described and because they are presented precisely in our earlier articles (Juhola et al., 2009, 2011, 2013).

2. The data set

Our data set included 19 variables that were computed for nystagmic beat candidates in order to advance their identification into accepted and rejected beats. Not all of these variables are medically useful, i.e., it is probable that otoneurological disorders do not affect all of them. In our previous research we, however, observed that all 19 variables are useful for the present classification task (Juhola et al., 2013). The variables are given in Table 1. See also Fig. 2 for symbols. The variable distributions are given in detail in (Juhola et al., 2013).

The data set included one signal from each of 107 patients suffering chiefly from acute, unilateral, peripheral loss of vestibular function, in other words, vestibular neuritis or having had a surgery for acoustic neuroma. Nystagmus of a patient sitting in a chair in the darkened room was measured with two small eye movement video cameras that detected eye movements with an image processing system. Each signal was 30 s long and included 20–80 acceptable nystagmic beats.

Every signal was analysed with a nystagmus detection algorithm (Juhola et al., 2011) and nystagmic beat candidates were, at the same time, separated either into accepted or rejected beats according to our algorithm (Juhola et al., 2013). Independent of this automatic separation, an expert manually explored all nystagmic

Table 1 Variables of the nystagmus data set defined with symbols associated with Fig. 2. In the variable names, *sp* and *fp* are slow and fast phases of nystagmus, *a* denotes amplitude, *d* duration, *v* mean velocity, *c* correlation, *mv* maximum velocity, *q* quality signal of the torsional component, and subscripts *h*, *v* and *t* indicate horizontal, vertical and torsional component signals.

Variables	Explanation	Definition
spa_h, spa_v, spa_t	Amplitudes of slow phase	$\lvert x(f) - x(s)\rvert$
fpa_h, fpa_v, fpa_t	Amplitudes of fast phase	$\lvert x(e) - x(f)\rvert$
spd_h	Duration of slow phase	$(f - s)/fr$, fr sampling frequency
fpd_h	Duration of fast phase	$(e - f)/fr$, fr sampling frequency
spv_h, spv_v, spv_t	Mean velocities of slow phase	Slope according to linear regression from samples $\{x(s), x(s+1), \ldots, x(f)\}$
fpv_h, fpv_v, fpv_t	Mean velocity of fast phase	Slope according to linear regression from samples $\{x(f), x(f+1), \ldots, x(e)\}$
spc_h, spc_v, spc_t	Correlations of slow phase	Correlation coefficient between samples $\{x(s), x(s+1), \ldots, x(f)\}$ and $\{y(s), y(s+1), \ldots, y(f)\}$
$fpmv_h$	Maximum velocity of horizontal fast phase	Slopes $z(i)$ according to linear regression from samples $\{x(i-2), x(f-1), x(i), x(i+1), x(i+2)\}$, $i = f+2, \ldots, e-2$, and then by taking $\max_{i \in \{f+2, \ldots, e-2\}} \{z(i)\}$
q_t	Mean of torsional signal quality values	Mean of quality values during slow phase of interval $[s, f]$

beat candidates, signal by signal, accepting or rejecting nystagmic beats based on visual screening. The purpose of applying both manual and automatic ways was to compare their results and also to use both of them jointly to see whether classification results could further be improved this way.

3. Data cleaning procedures

3.1. Straightforward cleaning procedure

In our original cleaning procedure (Juhola et al., 2013) we assumed that there would be two classes in the data having their class centres in different areas. This assumption was reasonable in the sense that the poor nystagmic beat candidates were marked to be rejected (manually or automatically) typically since some of their variable values were above some upper bounds. For instance, segments (2)–(5) in Fig. 3 would create too high torsional mean velocities for slow or fast phases of nystagmus because of the very steep spikes in the torsional signal. Cleaning was made first by computing the class centres and second by deleting those rejected nystagmic beats that were the closest to the class centre of the opposite class until the class size of the rejected nystagmic beat candidates became as small as the opposite class.

A more precise data analysis showed, however, that the class distribution was complicated so that no two clearly separate class centres appeared. The distributions computed with principal component analysis are given after the automatic rejection and acceptance of nystagmic beat candidates in Fig. 4(a) and similarly in Fig. 4(b) after first reducing the larger class of the rejected candidates according

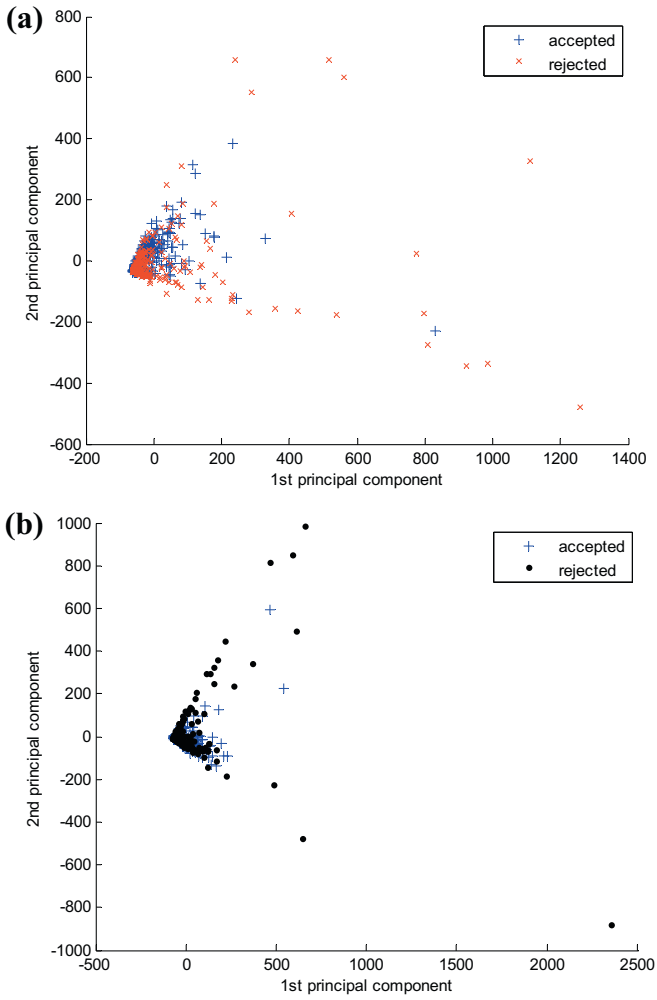

Fig. 4 (a) Before neighbourhood cleaning and (b) after this. The two first (most important) principal components were used to show the distributions of the accepted and rejected nystagmic beats after the automatic selection to the accepted and rejected beats. To make the scatter plots clear, one tenth of all beats only were drawn since their occurrences overlap considerably, particularly in the "left corner" of either distribution. The two first principal components accounted for 80% of variance in (a) before cleaning and 87% in (b) after cleaning.

to the cleaning procedure called neighbourhood cleaning (Laurikkala, 2001) described below. Fig. 4(a) shows how heavily the distributions of two classes were overlapping and that how cleaning given in Fig. 4(b) alleviated this difficulty.

After the manual selection of nystagmic beat candidates there were 2171 accepted and 3818 rejected beats. After the automatic selection these numbers were 2517 and 3472, and after using the both ways jointly 1645 and 4344, respectively. Thus, after cleaning, i.e., reducing the larger class of the rejected beats, the numbers were 2171, 2517 and 1645 for each of the classes in these three situations.

3.2. Neighbourhood cleaning procedure

The above straightforward cleaning procedure even impaired the classification accuracies from 80–90% (without cleaning) down to half for nearest neighbour searching, naïve Bayes rule and logistic discriminant analysis (Juhola et al., 2013), whereas linear and quadratic discriminant analysis and support vector machines with the linear and quadratic kernels obtained minor improvements. For the sake of these unexpected negative effects we abandoned the preceding cleaning procedure in the present study and used another, more sophisticated method (Laurikkala, 2001) that was based on nearest neighbour searching to determine which rejected beat candidates were the best to drop out from that larger class in order to balance the class sizes and, most of all, to use cleaning in order to improve classification.

According to (Laurikkala, 2001) the class of the rejected nystagmic beats was reduced by leaving out such rejected beat candidates that were found on the basis of two consecutively executed rules.

3.2.1. Neighbourhood cleaning procedure

(1) Iterate every nystagmic eye movement candidate c one by one as follows.
(2) Compute $k = 3$ nearest neighbours of c along with the Euclidean distance throughout n beats of the whole data set.
(3) If candidate c was from the class of the rejected beats, rule (3.1.) is tested. Otherwise, it was from the class of the accepted nystagmic beats and then rule (3.2.) is tested.
    (3.1.) If majority, in this case either 2 or 3 nearest neighbours were from the opposite class, accepted beats, the current candidate c is marked to be removed from the data set.
    (3.2.) Correspondingly, if the majority of k nearest neighbours were from the opposite class, rejected beats, these 2 or 3 rejected beats are marked to be removed from the data set.
(4) All those selected to be removed are left out from the data set.

The cleaning procedure acts locally, not globally as the previous that was constructed on the basis of the class centres. Thus, the neighbourhood cleaning method of (Laurikkala, 2001) is not hampered by a complicated distribution of more or less overlapping, mixing classes or one class surrounding the other. Since the cleaning procedure processes data cases on the basis of their nearest neighbourhood approach, it functions locally, independent of the "global" properties of data. It also cleans not only according to the majority class of the rejected beats, but also by means of the minority class. Nevertheless, in our nystagmus data cleaning is only directed to the majority class. This is very natural particularly

for the current data set, because the majority class represents rejected nystagmic eye movement beats and the aim was also to enhance dissimilarity between the classes of the accepted and rejected beats in order to improve classification results.

Note that the present cleaning procedure when directed to the majority class does not necessarily balance much the class sizes, but the number of removed items depends on the data. On the other hand, since small $k = 3$ was used, the two cleaning rules of the method explore tiny areas from the distribution space at a time and may, therefore, mark relatively many items to be removed, because the majority condition may be frequently satisfied.

3.3. Artificial extension of training data

Not pure cleaning is the only useful data refinement technique. Sometimes the use of artificial extension of data cases may also be useful, especially for balancing a class distribution (Swingler, 1996; Autio et al., 2007). In our data, class balancing was not necessary, because there were only two classes with frequent nystagmic beat candidates. Nevertheless, we experimented with the artificial extension by extending the same number of artificial nystagmic beats in the class of the rejected as was cleaned out. Such artificial cases should naturally resemble very much those of the same class. A simple way is just to copy existing cases. However, this is not perhaps fully suitable for all classification methods that may take advantage of the property that all cases are more or less different, i.e., multiple cases do not bring new information for training a model. Therefore, we created artificial rejected nystagmic beat candidates by calculating means of pairs of two successive rejected beats in the data set. (After cleaning, more nystagmic beat candidates still remained than were discarded from the class of the rejected beats.) We may assume that since two successive rejected beats are temporarily quite close to each other in a time-dependent signal, their nystagmus variable values may resemble each other. Consequently, computing the means of all their variable value pairs we may assume that such an average, artificial case would also be in the class of the rejected beats if it were real. Of course, several, more complicated, but perhaps more productive extension ways could be designed. We used this simple way because we assumed whatsoever that artificial extension could not improve results essentially since classification results obtained after neighbourhood cleaning were already very high.

It is important to know that the artificial cases were only used to extend training sets. It would not be sensible to include them in test sets.

4. Results

4.1. Classification procedure

As mentioned, the data set was computed according to three selections (Juhola et al., 2013): manual selection to accepted or rejected nystagmic beats, automatic

selection to the preceding two classes and both selections jointly. In addition, the neighbourhood cleaning procedure was run after each of the three selections.

In Fig. 5, there is a diagram how classifications were performed. To clarify the following descriptions of tests made, there is a round-up for the classification runs in Fig. 6.

We tested with nearest neighbour searching, Naïve Bayes rule, and linear, quadratic and logistic discriminant analysis, and support vector machines with different kernel functions. Results are presented in percents as classification accuracies, i.e., ratios of true positive added with true negative rates to the number of all test cases. Since the leave-one-out method was applied to tests, the number of tested cases was equal to all nystagmic beat candidates, 5989, or less when the data were cleaned.

Nearest neighbour searching and logistic discriminant analysis gave better classification results when the data were first standardized by subtracting the mean of each variable and then by dividing with its standard deviation. For those other classification methods, standardization (being not useful) was not used. Without standardization, the cleaning procedure removed 1313, 1680 and 1364 nystagmic beat candidates from the class of the rejected, respectively, after the three selections. With standardization, these were only 998, 717 and 732. Consequently, the class of the rejected beats remained as the majority class for all other situations than that without standardization for the automatic selection.

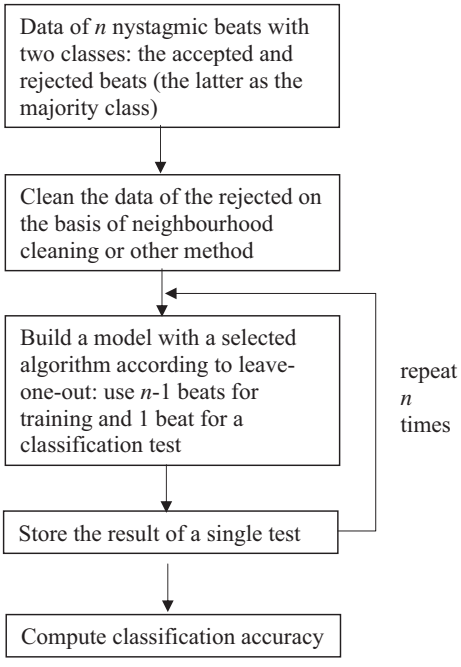

Fig. 5 The diagram of the classification procedure.

(1) Classification: Nearest neighbour searching with k=15 (results in Table 2)

no cleaning	random cleaning (1/4 out of the rejected)	random cleaning (1/2 out of the rejected)	neighbourhood cleaning (k=3)	neighbourhood cleaning (k=3) followed by artificial extension

(2) Classification: Linear (LD), quadratic (QD) and logistic discriminant (LogD) analysis, and naïve Bayes (NB) rule without or with neighbourhood cleaning with k=3 (results in Table 3)

LD, no cleaning	LD, cleaning	QD, no cleaning	QD, cleaning	LogD, no cleaning	LogD, cleaning	NB, no cleaning	NB, cleaning

(3) Classification: Support vector machines with linear (L), quadratic (Q) and radial basis function (RBF) kernels without or with neighbourhood cleaning with k=3 (results in Table 4)

L, no cleaning	L, cleaning	Q, no cleaning	Q, cleaning	RBF, no cleaning	RBF, cleaning

(4) Classification: Nearest neighbour searching (k=15) after running neighbourhood cleaning procedure with different k values (results in Table 5)

k equal to	5	7	9	11	13

Fig. 6 The round-up of all classification alternatives accomplished.

4.2. Classification with nearest neighbour searching

In Table 2, the results produced with k nearest neighbour searching are shown. The nearest neighbour number of k equal to 15 was used here since earlier it gave us at least 1% better results than odd values of less than 10 or greater than 20 (Juhola et al., 2013). To experimentally study the influence of the neighbourhood cleaning procedure (Laurikkala, 2001) described above, we also ran cleaning by

Table 2 Accuracy results of $k = 15$ nearest neighbour searching after data standardization and different cleaning procedures when the sizes of the classes of the accepted nystagmic beats were 2171, 2517 and 1645 for the three methods of the nystagmic beat selection. The accuracies of the best alternatives were marked in bold.

Cleaning procedure		Selection method of acceptance or rejection		
		Manual	Automatic	Manual and automatic
No cleaning	Class size of the rejected	3818	3472	4344
	Accuracy%	84.2	89.3	88.4
Random cleaning (quarter)	Class size of the rejected	2864	2604	3258
	Accuracy%	84.0	89.1	87.9
Random cleaning (half)	Class size of the rejected	1909	1736	2172
	Accuracy%	80.0	89.6	86.6
Neighbourhood cleaning	Class size of the rejected	2820	2755	3612
	Accuracy%	**91.7**	**94.9**	**94.4**
Neighbourhood cleaning and extension	Class size of the rejected	2820	2755	3612
	Accuracy%	**91.9**	**95.2**	**94.7**

leaving out a half or quarter from the class of the rejected beats as systematically dropping out every two or four beat candidates from those of the rejected. Let us call these 'random cleaning' because no assessment criterion for cleaning was employed. The last row in Table 2 includes the results when, after neighbourhood cleaning, artificial beats as many as cleaned beats from the rejected were inserted into the class of the rejected.

According to Table 2, we see that it is quite useless to run random cleaning, at least for this complicated data distribution, since the results were not better than those original of the uppermost row. (Random cleaning was made only to show that more efficient cleaning is necessary.) When strong cleaning that left the half out of the rejected was run, we even obtained partially poorer results. Of course, we also have to notice the different class sizes here that affect a priori probabilities in classification. Doubtless did neighbourhood cleaning affect positively by increasing classification accuracies by 5–6% for all three selection ways of nystagmic beat candidates. Instead, the extension of artificial, rejected beats after neighbourhood cleaning had only a minor effect, less than 0.5%. In principle, it might sometimes even impair results because there is no guarantee that such extension would always ''improve the quality of data''. Therefore, it is not reasonable to apply for tests of the subsequent tables.

4.3. Classification with discriminant analysis and Naïve Bayes rule

In Table 3 there are results of linear, quadratic and logistic discriminant analysis and Naïve Bayes rule. The different sizes of the classes of the rejected beats for the different classification methods came from whether the data standardization was used or was not used. Again neighbourhood cleaning was effective improving accuracies by 4–10% compared to the results of not cleaned situations.

Table 3 Accuracy results of discriminant analysis (with data standardization for the logistic one) and Naïve Bayes rule when the sizes of the classes of the accepted nystagmic beats were 2171, 2517 and 1645 for the three selection methods. The accuracies of the best alternative were marked in bold.

Classification method	Class size or accuracy	Selection method of acceptance or rejection		
		Manual	Automatic	Manual and automatic
Linear discr. anal., no cleaning	Rejected	3818	3472	4344
	Accuracy%	75.9	83.7	79.2
Linear discr. anal. with cleaning	Rejected	2505	1792	2980
	Accuracy%	83.7	89.6	83.9
Quadratic discr. anal., no cleaning	Rejected	3818	3472	4344
	Accuracy%	63.4	75.9	72.4
Quadratic discr. anal. with cleaning	Rejected	2505	1792	2980
	Accuracy%	80.3	86.0	79.8
Logistic discr. anal., no cleaning	Rejected	3818	3472	4344
	Accuracy%	80.5	87.2	86.9
Logistic discr. anal. with cleaning	Rejected	2820	2755	3612
	Accuracy%	**85.4**	**92.7**	**91.3**
Naïve Bayes rule, no cleaning	Rejected	3818	3472	4344
	Accuracy%	63.3	76.8	74.3
Naïve Bayes rule with cleaning	Rejected	2505	1792	2980
	Accuracy%	70.4	86.1	80.2

Table 4 Accuracy results of support vector machines without and with neighbourhood cleaning when the sizes of the classes of the accepted nystagmic beats were 2171, 2517 and 1645 for the three selection methods. Three kernel functions were used. The accuracies of the best alternatives were marked in bold.

Kernel, no cleaning or with it	Class size or accuracy	Selection method of acceptance or rejection		
		Manual	Automatic	Manual and automatic
Linear kernel, no cleaning	Rejected	3818	3472	4344
	Accuracy%	75.6	83.8	79.3
Linear kernel with cleaning	Rejected	2505	1792	2980
	Accuracy%	83.7	89.7	84.0
Quadratic kernel, no cleaning	Rejected	3818	3472	4344
	Accuracy%	81.9	88.4	85.5
Quadratic kernel with cleaning	Rejected	2505	1792	2980
	Accuracy%	**88.0**	**92.6**	**89.6**
RBF kernel, no cleaning	Rejected	3818	3472	4344
	Accuracy%	63.8	57.9	72.5
RBF kernel with cleaning	Rejected	2505	1792	2980
	Accuracy%	**88.7**	**88.2**	**88.0**

4.4. Classification with support vector machines

We also experimented with support vector machines the results of which are given in Table 4. Suitable parameter values were extensively studied as in (Juhola et al., 2013) and those giving the best average accuracy results were chosen for the three selection methods of nystagmic beat candidates (three rightmost columns in Table 4): (1) box constraint 8.9 for linear kernel and 0.1 for quadratic kernel, and box

constraint 1.2 and sigma 9.9 for radial basis function (RBF) kernel; (2) box constraint 1.2 for linear kernel and 1.0 for quadratic kernel, and box constraint 8.0 and sigma 10.0 for RBF kernel; (3) box constraint 0.2 for linear kernel and 0.1 quadratic kernel, and box constraint 0.7 and sigma 10.0 for RBF kernel; Of the three kernels, the quadratic kernel yielded 0–5% better accuracies than the other two. For RBF kernel, cleaning was essential, because without cleaning it lost all test cases stemming from the class of the accepted nystagmic beats and was even 25% worse than with cleaning. With cleaning it was 4–5% worse than the other two kernels.

4.5. Classification with nearest neighbour searching after neighbourhood cleaning with different k values

Ultimately, we still studied the use of greater k values than 3 for the neighbourhood cleaning procedure. Results obtained are given in Table 5. These are only shown for the classification set-up where nearest neighbour searching was applied to classification since similar phenomena were also expected for the other classification methods. Increasing the number up to 11 or 13 of the nearest neighbours searched for in the cleaning procedure improved classification accuracies slightly, by 1–2%. This came from the stronger cleaning, in other words, with greater k values more elements were left out from the class of the rejected beats. Therefore, the more intensive data cleaning may be a reasonable approach if there are abundant elements in the majority class as were here.

It was obvious that increasing k in neighbourhood cleaning did not much improve the results obtained compared to those of k equal to 3, because the latter were already high, over 90%, although more and more elements of the majority class were left out. Note two lowest rows in Table 5. According to them, it seemed in the present data that increasing k over 11 did not improve classification accuracies.

Table 5 Classification accuracy results of $k = 15$ nearest neighbour searching after the neighbourhood cleaning procedure when the sizes of the classes of the accepted nystagmic beats were 2171, 2517 and 1645 for the three methods of the nystagmic beat selection and when in cleaning 5, 7, 9, 11 or 13 nearest neighbours were utilized to determine elements to be cleaned. The accuracies of the best alternatives were marked in bold.

Neighbours in cleaning		Selection method of acceptance or rejection		
		Manual	Automatic	Manual and automatic
5	Class size of the rejected	2607	2633	3469
	Accuracy%	92.4	95.8	95.0
7	Class size of the rejected	2455	2536	3365
	Accuracy%	92.8	96.3	95.3
9	Class size of the rejected	2325	2467	3288
	Accuracy%	93.1	96.4	95.6
11	Class size of the rejected	2184	2407	3196
	Accuracy%	**94.1**	**97.0**	**96.0**
13	Class size of the rejected	2086	2362	3113
	Accuracy%	**94.2**	**96.9**	**96.0**

5. Discussion and conclusions

In all of our results, neighbourhood cleaning was an efficient way to refine data and reduced the size of the majority class, rejected nystagmic beats. Simpler cleaning ways could hardly function as effectively, at least neither that was used earlier (Juhola et al., 2013) nor in Table 2.

Logistic discriminant analysis after cleaning was 2–12% better than the others in Table 3 and 2–6% worse than the accuracies of the nearest neighbour searching after neighbourhood cleaning as in Table 2. However, the best support vector machines (with the quadratic kernels) were 2–5% poorer than the nearest neighbour accuracies after neighbourhood cleaning as in Table 2. This is a slightly surprising conclusion since our experience among some other data sets has been the opposite, frequently support vector machines have been subtly better than others. We assume that the possible reason was now that neighbourhood cleaning favoured the nearest neighbour searching classification. Since the nearest neighbour searching was applied in both, cleaning could ''favour'' its ''relative classification method'' more than the others.

The automatic selection was better in some cases than the manual and automatic ones together. Apparently, this stemmed from the fact that manual selection criteria may vary a little from time to time. Instead, the automatic selection always functions stably.

The use of greater nearest neighbour numbers (5, 7, 9 or 11) than 3 originally used improved the classification results slightly further since more elements were cleaned out from the majority class compared to the situation of 3 nearest neighbours. However, no such conclusion could be drawn that this phenomenon would typically be present. After all, the properties of data and their distribution are essential.

A future research detail in using the neighbourhood cleaning method could be to attempt to also clean the minority class. In general this is not perhaps sensible, but for such data sets as ours here where both classes were fairly large, it might be useful since the classes were ''mutually overlapping'', some rejected nystagmic beats among the accepted and vice versa, as seen in Fig. 4. On the other hand, cleaning accepted beats should be made very carefully and ''conservatively'', not to deteriorate the validity of the data set as a training set for nystagmus analysis viz., the accepted nystagmic beats represent physiologically plausible and possible nystagmus variable values, whereas the rejected beats are more or less values outside physiologically possible boundaries. However, these boundaries are not exact, because they may vary between subjects.

We can conclude that neighbourhood cleaning refined data efficiently and improved accuracies throughout the tests accomplished. Its character is rather local than global and it can purify noise-like occurrences from data. It had also good influence on the current data set with the complicated distribution. To clean data sets with unknown distributions, it is best to first explore their data distributions

– as this is useful in general – before cleaning data and to choose a cleaning procedure carefully.

Acknowledgements

The second author is thankful to Maj and Tor Nessling Foundation for the support.

References

Abel, L.A., Wang, Z.I., Dell'Osso, L.F., 2008. Wavelet analysis in infantile nystagmus syndrome. Invest. Ophthalmol. Vis. Sci. 49 (8), 3413–3423.

Augustyniak, P., 1996. Advanced method of nystagmus-phase separation using adaptive modification of time-frequency signal representation. In: 3rd International Conference on Signal Processing (ICSP'96), Beijing, China. pp. 351–354.

Autio, L., Juhola, M., Laurikkala, J., 2007. On the neural network classification of medical data and an endeavour to balance non-uniform data sets with artificial data extension. Comp. Biol. Med. 37, 388–397.

Hertle, R.W., Dell'Osso, L.F., 1999. Clinical and ocular motor analysis of congenital nystagmus in infancy. J. Am. Assoc. Pediatr. Ophthalmol. Strabismus 3 (2), 70–79.

Hosokawa, M., Hasebe, S., Ohtsuki, H., Tsuchida, Y., 2004. Time-frequency analysis of electronystagmogram signals in patients with congenital nystagmus. Jpn. J. Ophthalmol. 48, 262–267.

Juhola, M., 1988. Detection of nystagmus eye movements using a recursive digital filter. IEEE Trans. Biomed. Eng. 35 (5), 389–394.

Juhola, M., Aalto, H., Hirvonen, T., 2009. On signal analysis of three-dimensional nystagmus. In: Adlassning, K.-P., Blobel, B., Mantas, J., Masic, I. (Eds.), Proceedings of Medical Informatics in Europe 2009 (MIE2009), Sarajevo, Bosnia and Herzegovina. IOS Press, pp. 846–850.

Juhola, M., Aalto, H., Jutila, T., Hirvonen, T., 2011. Signal analysis of three-dimensional nystagmus for otoneurological investigations. Ann. Biomed. Eng. 39 (3), 973–982.

Juhola, M., Aalto, H., Joutsijoki, H., Hirvonen, T.P., 2013. The classification of valid and invalid beats of three-dimensional nystagmus eye movement signals using machine learning methods. Adv. Artif. Neural Syst.. http://dx.doi.org/10.1155/2013/972412. Article ID 972412.

Laurikkala, J., 2001. Improving identification of difficult small classes by balancing class distribution. In: Quaglini, S., Barahona, P., Andreassen, S. (Eds.), . In: Artificial Intelligence in Medicine: Eight European Conference on Artificial Intelligence in Medicine in Europe, Lecture Notes in Artificial Intelligence, vol. 2101. Springer, Berlin, pp. 63–66.

Swingler, K., 1996. Applying Neural Networks, A Practical Guide. Academic Press, London.

Tominaga, S., Tanaka, T., 2010. 3-Dimensional analysis of nystagmus using video and image processing. In: The Society of Instrument and Control Engineers, SICE Ann. Conf., Taipei, Taiwan. pp. 89–91.

Wall III, C., Black, F.O., 1982. Algorithms for the clinical analysis of nystagmus eye movements. IEEE Trans. Biomed. Eng. 28, 638–646.