Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 34 (2008) 628–634
Expert Systems with Applications
www.elsevier.com/locate/eswa

Visualization and analysis of classifiers performance in multi-class medical data

Banu Diri [1], Songul Albayrak *

Yildiz Technical University, Computer Engineering Department, A-Block, Floor-1, 34349 Istanbul, Turkey

Abstract

The primary role of the thyroid gland is to help regulation of the body's metabolism. The correct diagnosis of thyroid dysfunctions is very important and early diagnosis is the key factor in its successful treatment. In this article, we used four different kinds of classifiers, namely Bayesian, *k*-NN, *k*-Means and 2-D SOM to classify the thyroid gland data set. The robustness of classifiers with regard to sampling variations is examined using a cross validation method and the performance of classifiers in medical diagnostic is visualized by using cobweb representation. The cobweb representation is the original contribution of this work to visualize the classifiers performance when the data have more than two classes. This representation is a newly used method to visualize the classifiers performance in medical diagnosis.

Keywords: Bayesian; *k*-NN; *k*-Means; 2-D SOM; Cross validation; Confusion matrix; ROC analysis; Cobweb representation; Thyroid gland data; Medical diagnosis

1. Introduction

The thyroid gland is one of the most important organs in the body as thyroid hormones are responsible for controlling metabolism. As a result, thyroid function impacts on every essential organ in the body. The thyroid gland is located in the front of the neck attached to the lower part of the voice box (or larynx) and to the upper part of the windpipe (or trachea). It weighs only about 25 g. It has two sides or lobes. These lobes are connected by a narrow neck (or isthmus). Each lobe is about 4 cm long and 1–2 cm wide. However, the hormones it secretes are essential to all growth and metabolism. The gland is a regulator of all body functions. The name "thyroid" comes from the Greek word, which means, "shield" (http://www.thyroid.ca/Guides/HG01.html).

The thyroid gland produces thyroid hormones. These are peptides containing iodine, which plays an important role in the function of the thyroid gland. The two most important hormones are tetraiodothyronine (thyroxine or T4) and triiodothyronine (T3). These hormones are essential for life and have many effects on body metabolism, growth, and development (http://www.thyroid.ca/Guides/HG01.html). Hormones produced by two other organs influence the thyroid gland. One of them is the pituitary gland, which located at the base of the brain produces thyroid stimulating hormone (TSH). The other one is the hypothalamus, which is a small part of the brain above the pituitary and produces thyrotropin releasing hormone (TRH).

The most common thyroid disorder is an under active thyroid, known as hypothyroidism, in which the thyroid does not produce enough hormone. Less frequently, the thyroid produces too much hormone, which is known as hyperthyroidism. The state of normal thyroid function is called euthyroidism (Zhang & Berardi, 1998). About 200 million people in the world have some form of thyroid

* Corresponding author. Tel./fax: +90 212 258 74 89.
E-mail addresses: banu@ce.yildiz.edu.tr (B. Diri), songul@ce.yildiz.edu.tr (S. Albayrak).
[1] Tel.: +90 212 327 36 73; fax: +90 212 258 74 89.

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.10.016

disease. Approximately 2–3% of the general population in the United States suffers from either hypothyroidism or hyperthyroidism (Tunbridge, 1997). One in twenty of the Canadian population is affected with abnormalities of the thyroid gland (http://www.thyroid.ca/Guides/HG01.html). All thyroid disorders are much more common in women than in men.

Hypothyroidism affects approximately 2 persons in 100. The signs and symptoms of overt hypothyroidism are opposite to those in hyperthyroidism since there is a deficiency of thyroid hormone secretion and all metabolic processes "slow down". The patient has poor appetite, intolerance to cold, dry, coarse, skin, brittle hair, tiredness, a croaky, hoarse voice, constipation, and muscle weakness. Examination may reveal an absence of the thyroid gland, dry, scaly, cold, pale skin, a thickening of the skin and underlying tissues, very slow reflexes and a slow heart rate. The patient can have poor memory retention. The diagnosis of hypothyroidism is confirmed by finding very low levels of thyroid hormones (T4 and T3) in the blood. The symptoms of hyperthyroidism vary from patient to patient, and may include one or more of the following: The patient has fast heart beats, shortness of breath, anxiety, difficulty in concentrating, feeling warm and/or increased sweating, muscle weakness, sleep disorders, fatigue or increased energy, hair loss, acne, oily skin, weight loss or less commonly weight gain, menstrual irregularities (http://www.mythyroid.com).

Thyroid disorders for the most part are treatable; however, untreated thyroid disease produces serious results in other parts of the body. The correct diagnosis of thyroid dysfunctions based on clinical and laboratory test often proves difficult. The difficulty in diagnosis comes from the inconsistency in test results across patients and other factors such as pregnancy, drug interactions, non thyroidal illnesses, and psychiatric problems which are all known to affect the thyroid hormone levels measured in the laboratory tests (Gavin, 1988; Wong & Steffes, 1984).

In this work, Bayesian and *k*-NN classification are used as supervised classification methods and *k*-Means and self-organizing feature map (SOFM) are used as unsupervised clustering methods on thyroid gland data obtained by Coomans, Broeckaert, Jonckheer, and Massart (1983). Research activities have shown that statistical methods, machine learning algorithms and artificial neural networks (ANNs) have powerful medical prediction ability to achieve accurate automatic diagnosis. They have been used extensively in many different problems including thyroid function diagnosis. In a very recent work, Ozyilmaz and Yildirim investigate Dr. Coomans's thyroid gland data on the supervised classification methods to develop a medical diagnostic system. Their work includes only the neural network classification methods such as MLP, RBF and adaptive CSFNN networks (Ozyilmaz & Yildirim, 2002). Zhang and Berardi (1998) have also investigated neural network methods on another thyroid gland data, which is highly unbalanced data set. Their results showed that neural networks provide good classification rate for smaller group members. Thyroid gland data set was used with unsupervised clustering methods by Albayrak (2003). However, the work presented here covers a more extensive study than the previous one about classification methods for thyroid gland data set and furthermore the cobweb representation is the original part of the study to visualize the classifiers performances.

Table 1
Confusion Matrix

Predicted	Actual	
	T	F
T	True Positives (TP)	False Positives (FP)
F	False Negatives (FN)	True Negatives (TN)

In general, classifiers are used to make predictions for decision support. Since predictions can be wrong, it is important to know what the effect is when the predictions are incorrect. In many situations not every error has the same consequences. Some errors end up in greater medical expenditure than others, especially in medical diagnosis. For instance, a wrong diagnosis or treatment can result in severe expenditure and dangers depending on the kind of mistake that has been done (Ferri, Hernandez-Orallo, & Salido, 2003). Consequently, accuracy is not generally the best way to evaluate the quality of a classifier or a learning algorithm.

The Receiver Operation Characteristic (ROC) analysis allows the evaluation of a classifier performance in a more independent and complete way than just using accuracy (Provost et al., 1997; Sweets, Dawes, & Monahan, 2000). ROC analysis has usually been presented for two classes, because it is easy to define and interpret and it is computationally feasible. In ROC analysis with two classes, the following notation is used for the *confusion matrix* (Table 1).

In thyroid gland data set, there are three classes and a *class confusion matrix* is useful to analyze errors in the case of $k > 2$ classes (k represents the number of classes). It is a $(k \times k)$ matrix such that its entry (i,j) contains the number of instances that belongs to C_j (class j) but is assigned to C_i (class i). Ideally all off-diagonals should be 0, for no misclassification. The class confusion matrix allows us to pinpoint what types of misclassification occur (Alpaydın, 2004).

The paper is organized as follows. The next section contains background about supervised and unsupervised classification methods. Information about cross validation method and thyroid gland data set are given in Sections 3 and 4, respectively. The results are presented in Section 5. The final section contains conclusions.

2. Background

2.1. Supervised methods

There are different methods as supervised classification. Supervised classification requires a priori knowledge of the number of classes, as well as knowledge concerning statistical aspects of the classes. All methods start with establishing training samples, which are areas that are assumed or verified to be of a particular type. In this work, we use 3-fold cross validation to obtain training set and Bayesian and k-Nearest Neighbor classification algorithms as supervised methods.

Bayesian Classifier. Bayesian decision theory is fundamental statistical approach to the problem of pattern classification. This approach is based on the assumption that the decision problem is posed in probabilistic terms, and that all of the relevant probability values are known. Bayesian classification yields an optimum classifier when the probability density function of each pattern population and the probability of occurrence of each pattern class are known.

k-Nearest Neighbor Classifier. The k-NN classifier assigns the input to the class having most examples among the k neighbors of the input. All neighbors have the equal vote, and the class having the maximum number of voters among the k neighbors is chosen. Ties are broken arbitrarily or a weighted vote is taken. k is generally taken an odd number to minimize ties (Alpaydın, 2004). In this study, we choose k as 1 to classify the thyroid gland data.

2.2. Unsupervised methods

When performing an unsupervised classification, it is necessary to find the right number of classes that are to be found. For this application, we know the number of classes because we use thyroid gland data, which has three classes as hypothyroid, hyperthyroid, and euthyroid.

K-Means. In k-Means clustering algorithm which is a non-hierarchical approach to form exact clusters, a set of n data points are given in d-dimensional space and an integer k, and the problem is to determine a set of k points to minimize the mean squared distance from each data point to its nearest center. One disadvantage is that it is a local search procedure and the final cluster centers highly depend on the initial cluster centers (Alpaydın, 2004). We prefer to select initial cluster centers from the data by dividing each attribution into k equal intervals instead of random selection.

Self-organizing feature map (SOFM). Self-organizing mapping is a kind of neural network, which is based on competitive learning. The output neurons of the network compete among themselves to be activated, with the result that only one output neuron or one neuron per group wins the competition. The output neurons that win the competition are called winner-take-all neurons. One way of inducing winner-take-all competition among the output neurons is to use lateral inhibitory connections between them (Bose & Liang, 1996; Haykin, 1994; Schalkoff, 1992).

In a self-organizing feature map, the neurons are placed at the nodes of a lattice that is usually one or two-dimensional. The neurons become selectively tuned to various input patterns in the course of a competitive learning process. The location of the neurons (i.e. winning neuron) so tuned tends to become ordered with respect to each other in such a way that a meaningful coordinate system for different input feature is created over the lattice (Haykin, 1994).

The learning process involved in the computation of a feature map is stochastic in nature, which means that the accuracy of the map depends on the number of iterations of the SOFM algorithm. Moreover, the success of map formation is critically dependent on how the main parameters of the algorithm, namely, the learning rate parameter η and the neighborhood function A_i, are selected (Haykin, 1994). In this application, we use learning rate parameter as $\eta(t) = 1/\sqrt{t}$ in 2-D SOFM algorithm with the 5×5, 6×6, 7×7 and 8×8 nodes of lattice.

3. Cross validation

Cross validation can be used simply to estimate the generalization error of a given model, or it can be used for model selection by choosing one of several models that has the smallest estimated generalization error. The holdout method is the simplest kind of cross validation. The data set is separated into two sets, called the training set and the testing set. The function approximator fits a function using the training set only. Then the function approximator is asked to predict the output values for the data in the testing set. The errors occurred are accumulated as before to give the mean absolute test set error, which is used to evaluate the model. The advantage of this method is that it is usually preferable to the residual method and takes no longer to compute. However, its evaluation can have a high variance. The evaluation may depend heavily on which data points end up in the training set and which end up in the test set, and thus the evaluation may be significantly different depending on how the division is made (http://www.faqs.org/faqs/ai-faq/neural-nets). K-fold cross validation is one way to improve over the holdout method. The data set is divided into k parts, and the holdout method is repeated k times. Each time, one of the k parts is used as the test set and the other $k - 1$ parts are put together to form a training set. Then the average error across all k trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set $k - 1$ times. The variance of the resulting estimate is reduced as k is increased.

In the work presented here, we have chosen a threefold cross validation method, therefore the data set for three different classes is divided into three mutually exclusive partitions of approximately equal size. Training is performed on two of the partitions while the third is used for the testing purpose. The process is repeated until each partition has served as the test data. The training sample is used for model fitting and/or parameter estimation.

Table 2
Attributes used for thyroid diagnosis

1 T3-resin uptake test (A percentage)
2 Total Serum thyroxin as measured by the isotopic displacement method
3 Total Serum triiodothyronine as measured by radioimmuno assay
4 Basal thyroid-stimulating hormone (TSH) as measured by radioimmuno assay
5 Maximal absolute difference of TSH value after injection of 200 μg of thyrotropin-releasing hormone as compared to the basal value

4. Thyroid gland data set

The data set used in this study comes from Coomans et al. (1983), which contains information related to thyroid function. This data set contains three classes and 215 samples. These are assigned to the values that correspond to the hyper-, hypo- and euthyroidism (normal) function of the thyroid gland. The problem is to determine whether a patient has normally functioning thyroid, an under-functioning thyroid (hypothyroid) or over-active functioning thyroid (hyperthyroid). There are 215 samples in the data set and the hyperthyroid class denotes 16.3% (35 samples) of the data points, the rate of hypothyroid class is 13.9% (30 samples) of the data points, while the remaining 69% (150 samples) are the euthyroidisim class. Five laboratory tests are used to try to predict whether a patient's thyroid belongs to the class of euthyroidisim, hypothyroidism or hyperthyroidism. The diagnosis (the class label) was based on a complete medical record, including anamnesis, scan, etc. For each of the 215 samples, there are five attributes, which are continuous variables, used to determine to which of the three classes the patient belongs. Table 2 lists these five attributes and their description.

5. Experimental results

In this paper, we have investigated the potential of supervised classification methods, Bayesian and *k*-NN, and unsupervised classification methods *k*-Means and self-organizing feature map (SOFM), on thyroid dysfunction diagnosis. In the application, 2-D SOFM with 5 × 5, 6 × 6, 7 × 7 and 8 × 8 nodes of lattice is formed but 7 × 7 nodes of lattice are preferred which have the best result. Thyroid disease identification is an important yet difficult task from both clinical diagnosis and statistical classification points of view. The large number of interrelated patient attributes as well as extremely unbalanced groups in the thyroid diagnosis problem complicates the relationship between these attributes and the patient's true group membership, which causes poor performance for traditional model-based statistical methods.

Thyroid gland data set contains three classes and 215 samples. 150 samples belong to euthyroid class, 35 samples are from hyperthyroid class and 30 samples from hypothyroid class. Using 3-fold cross validation method, the data set is divided into three parts as Part-a consisting of 71 samples, Part-b consisting of 72 samples, and Part-c having 72 samples. According to this, one of the parts was left outside and the others were used for training.

5.1. Accuracy of the classifiers

For subsample-1, Part-a and the Part-b were trained and Part-c was tested for all methods. For this test sample, classification accuracies were obtained as 95.77% for Bayesian and 91.55% for *k*-NN, while the classification accuracies were 92.95% for 2-D SOM and 84% for *k*-Means. For subsample-2, the Part-a and Part-c were trained and Part-b was tested for all methods. Classification accuracies for this subsample were 98.61% for Bayesian, 94.44% for *k*-NN, 91% for *k*-Means, and 87.50% for 2-D SOM. For subsample-3, Part-b and Part-c were trained and Part-a was tested for all methods. The obtained classification accuracies for this test sample were 95.83% for Bayesian, 91.67% for *k*-NN, 84.72% for 2-D SOM, and 72% for *k*-Means. Performance results for all subsamples are given in Table 3. It can be seen from the results that Bayesian achieves higher classification rates than *k*-NN in supervised classification methods. However, 2-D SOM gives better results than *k*-Means for subsample-1 and subsample-3, whereas *k*-Means gives higher classification rates than 2-D SOM for subsample-2 in unsupervised classification methods.

Table 4 shows the overall classification accuracy, which is obtained by averaging the classification accuracy of three

Table 3
3-Fold cross validation performance results of the test subsamples

Method	Measure	Hyper	Hypo	Euth	Overall
Subsample 1					
Bayesian	% Correct	100.00	80.00	98.00	95.77
	Correct #	11	8	49	68
k-NN	% Correct	81.82	60.00	100.00	91.55
	Correct #	9	6	50	65
k-Means	% Correct	63.64	60.00	94.00	84.00
	Correct #	7	6	47	60
2-D SOM	% Correct	81.82	70.00	100.00	92.95
	Correct #	9	7	50	66
Subsample 2					
Bayesian	% Correct	100.00	100.00	98.00	98.61
	Correct #	12	10	49	71
k-NN	% Correct	91.67	80.00	98.00	94.44
	Correct #	11	8	49	68
k-Means	% Correct	91.67	90.00	92.00	91.00
	Correct #	11	9	46	66
2-D SOM	% Correct	75.00	70.00	94.00	87.50
	Correct #	9	7	47	63
Subsample 3					
Bayesian	% Correct	100.00	80.00	98.00	95.83
	Correct #	12	8	49	69
k-NN	% Correct	83.33	80.00	96.00	91.67
	Correct #	10	8	48	66
k-Means	% Correct	58.33	80.00	74.00	72.00
	Correct #	7	8	37	52
2-D SOM	% Correct	66.67	80.00	90.00	84.72
	Correct #	8	8	45	61

Table 4
Overall classification accuracy of each classifier on thyroid gland data set

	Hyper	Hypo	Euth	Overall
Bayesian	100.00	86.67	98.00	96.74
k-NN	85.61	73.33	98.00	92.55
k-Means	71.21	76.67	86.67	82.33
2-D SOM	74.50	73.33	94.67	88.40

Table 5
The class confusion matrix of subsample-3

Predicted	Actual					
	Supervised					
	Bayesian			*k*-NN		
	Euth	Hyper	Hypo	Euth	Hyper	Hypo
Euth	49	0	2	48	2	2
Hyper	0	12	0	0	10	0
Hypo	1	0	8	2	0	8
Total	50	12	10	50	12	10
	Unsupervised					
	k-Means			2-D SOM		
	Euth	Hyper	Hypo	Euth	Hyper	Hypo
Euth	37	5	2	45	4	0
Hyper	13	7	0	2	8	2
Hypo	0	0	8	1	0	8
Total	50	12	10	48	12	10

subsamples. Bayesian classifier can achieve 96.74% classification rate which is higher than that in *k*-NN with 92.55% in supervised methods. 2-D SOM method can achieve 88.40% classification rate which is higher than the *k*-Means giving 82.33% in unsupervised methods. The difference between the overall classification rates is only 4.19% for supervised classifiers and 6.07% for unsupervised classifiers according to Table 4.

5.2. Visual representation of ROC analysis with more than two classes

Receiver Operating Characteristic (ROC) analysis is widely used to analyze the performance of two-class classifiers. ROC analysis shows the trade-off between the sensitivity and specificity as the decision threshold varies. Sensitivity is the fraction of positive cases that are correctly classified as positive and specificity is the fraction of negative cases that are correctly classified as negative. Advantages of ROC analysis include the fact that it explicitly considers the trade-offs in sensitivity and specificity, includes visualization methods, and has clearly interpretable summary metrics (Patel et al., 2005). Currently, there does not exist a widely accepted performance method similar to ROC analysis for *k*-class classifiers ($k > 2$).

In this work, we used cobweb representation to visualize the performance of 3-class classifiers on the thyroid gland data set. Graphical toolbox ('radar'-plot) of Microsoft Excel was used to produce the cobweb graph. The cobweb graphical performance representation provides a quick way to visualize classifier performance. In a 3-class classifier, 3×3 class confusion matrix is used to analyze errors. For a chance classification, the misclassification values of confusion ratio matrix are (0.33, 0.33, 0.33, 0.33, 0.33, 0.33). A polygon within the chance performance hexagon shows a better performance than chance classifier. Currently, there is no numeric metric associated with the cobweb graphical representation (Patel et al., 2005).

The number of correct classified and misclassified samples from subsample-3 is given in Table 5 and is used in the Cobweb representation. Class confusion matrix was 3 by 3 for thyroid gland data set with three classes and Table 5 can also be thought as the class confusion matrix for each classifier. For the test set of subsample-3, 50 samples belong to euthyroid, 12 samples belong to Hyperthroid and 10 samples belong to Hypothyroid. When the test set was clustered by 2-D SOM with the 7×7 nodes of lattice, 45 of 50 samples were clustered as euthyroid, 2 of 50 samples were clustered as Hyperthyroid and 1 of 50 samples was clustered as Hypothyroid which belong to euthyroid. However, there were two samples which were not assigned to a class in 2-D SOM classification method. Therefore, 48 samples were clustered totally and this is shown as shaded in Table 5.

Each column in the class confusion matrix was normalized by the total value of this class and the class confusion ratio matrix was obtained as shown in Table 6.

To form a cobweb representation, the misclassification values are chosen from confusion ratio matrix. Each corner of the hexagon represents the normalized misclassification value which comes from the off-diagonal of class confusion ratio matrix. The values chosen from a three-class classification confusion matrix form a six-dimensional point. In this application, there were three classes, euthyroid, hyperthyroid and hypothyroid, and the following points (euthyroid → hyperthyroid, euthyroid → hypothyroid, hyperthyroid → euthyroid, hyperthyroid → hypothyroid,

Table 6
The class confusion ratio matrix of subsample-3

Predicted	Actual					
	Supervised					
	Bayesian			*k*-NN		
	Euth	Hyper	Hypo	Euth	Hyper	Hypo
Euth	0.98	0	0.2	0.96	0.17	0.2
Hyper	0	1	0	0	0.83	0
Hypo	0.02	0	0.8	0.04	0	0.8
	Unsupervised					
	k-Means			2-D SOM		
	Euth	Hyper	Hypo	Euth	Hyper	Hypo
Euth	0.74	0.42	0.2	0.9	0.33	0
Hyper	0.26	0.58	0	0.04	0.67	0.2
Hypo	0	0	0.8	0.02	0	0.8

hypothyroid → euthyroid, hypothyroid → hyperthyroid) were obtained as misclassification values from the class confusion ratio matrix. Euthyroid → hyperthyroid corresponds to euthyroid class objects, which were misclassified as hyperthyroid class objects. A chance classification is shown in Figs. 1 and 2. The represented points are (0.33, 0.33, 0.33, 0.33, 0.33, 0.33). The misclassification rates of 0.33 show that when confronted with an object of type euth, the classifier would classify it as having en equal likelihood of being from any of the three classes, euth, hypo and hyper.

To explain the usage of cobweb representation, we can inspect the classification results of Bayesian and *k*-NN classifier. Cobweb representation of subsample-3 for supervised classification methods is shown in Fig. 1. The misclassification ratios for Bayes and *k*-NN classifiers are the same as 20% in hypothyroid class for subsample-3 (hypothyroid → euthyroid). This misclassification ratio is shown on the hypothyroid → euthyroid corner and the 0.2 scale of the hexagon.

The classification results are better than the chance classifier, since the misclassification ratios are inside of the chance performance hexagon.

Cobweb representation of subsample-3 for unsupervised classification methods is shown in Fig. 2. The unsupervised

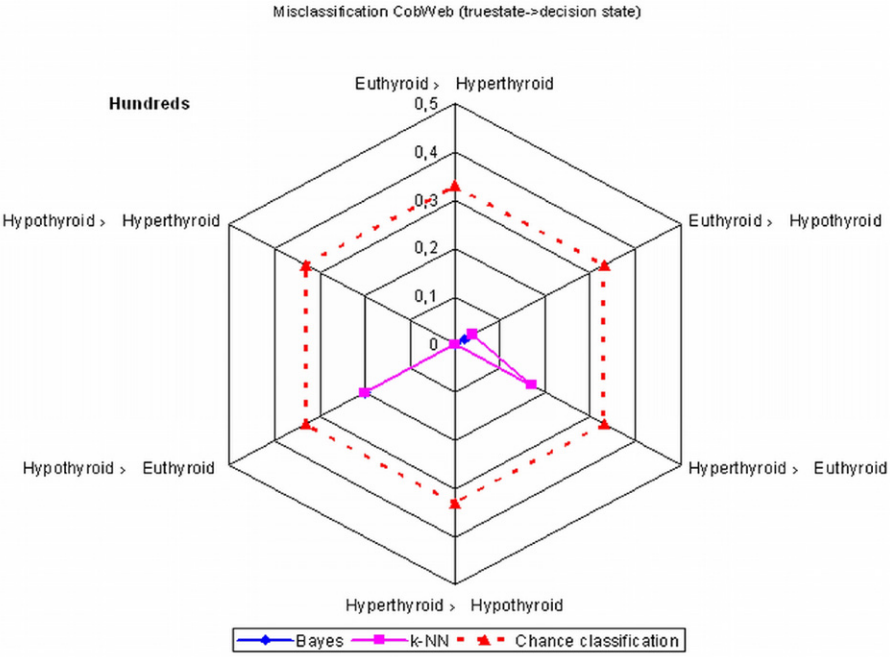

Fig. 1. Cobweb representation of subsample-3 for supervised methods.

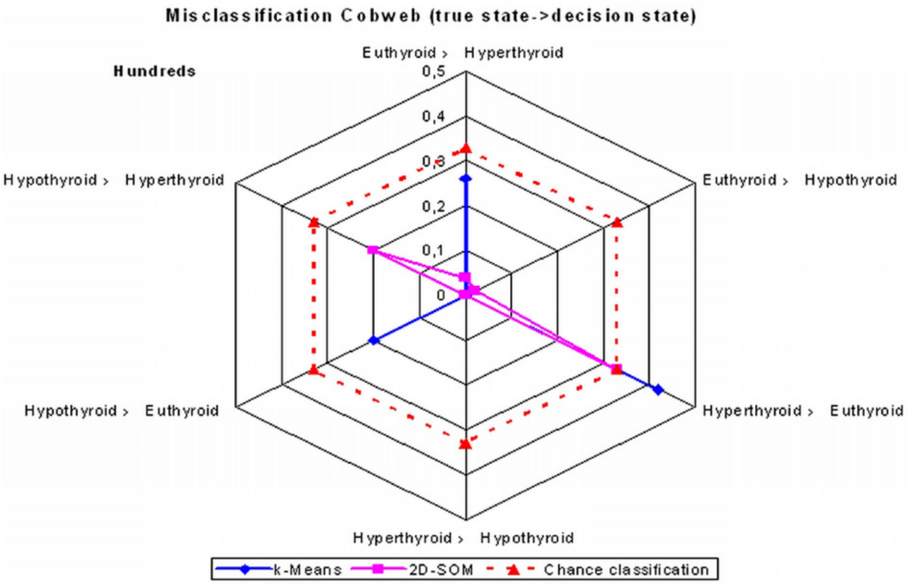

Fig. 2. Cobweb representation of subsample-3 for unsupervised methods.

classification methods are also better than the chance classifier for most cases except hyperthyroid → euthyroid for k-Means where it is worse than chance classifier and hyperthyroid → euthyroid for 2-D SOM where it is near to chance classifier.

6. Conclusions

In this paper, we have investigated the performance of classification algorithms on the diagnosis of thyroid dysfunctionality and visualized them in the form of multi-class extension of ROC analysis. Bayesian and k-NN classifications have been used as supervised classification methods and k-Means and 2-D SOM have been used as unsupervised clustering methods. We have coded all methods, which are used in this application, without using any toolbox. The robustness of classifiers with regard to sampling variations is examined using a cross validation method and the performance of classifiers in medical diagnostic is visualized by using cobweb representation. The cobweb representation is the original contribution of this work to visualize the classifiers performance when the data have more than two classes. This representation is a newly used method to visualize the classifiers performance in medical diagnosis. The performance of the classifiers has also been investigated visually and most of them have been found within the boundary of the chance hexagon.

References

Albayrak, S. (2003). Unsupervised clustering methods for medical data: an application to thyroid gland data. *Lecture Notes in Computer Science, 2714*, 695–701.

Alpaydın, E. (2004). *Introduction to machine learning*. The MIT Press.

Bose, N. K., & Liang, P. (1996). *Neural network fundamentals with graphs algorithms and applications*. Mc Graw Hill.

Coomans, D., Broeckaert, I., Jonckheer, M., & Massart, D. L. (1983). Comparison of multivariate discrimination techniques for clinical data—application to the thyroid functional state. *Methods of Information in Medicine, 22*, 93–101.

Ferri, C., Hernandez-Orallo, J., Salido, M. A. (2003). Volume under the ROC surface for multi-class problems. Exact computation and evaluation of approximations. Technical Report.

Gavin, L. A. (1988). The diagnostic dilemmas of hyperthyroxinemia and hypothyroxinemia. *Advances in International Medicine, 33*, 185–204.

Haykin, S. (1994). *Neural networks: a comprehensive foundation*. Macmillan College Publishing Company.

Ozyilmaz, L., Yildirim, T. (2002). Diagnosis of thyroid disease using artificial neural network methods. In *Proceedings of the ninth international conference on neural information processing (ICONIP'02)* (pp. 2033–2036).

Patel, A. C., Markey, M. K. (2005). Comparison of three-class classification performance metrics: a case study in breast cancer CAD. Medical Imagining 2005: Image Processing.

Provost, F., Fawcett, T. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distribution. In *Proceedings of the third international conference on knowledge discovery and data mining (KDD-97)* (pp. 43–48).

Schalkoff, R. J. (1992). *Pattern recognition: statistical, structural and neural approaches*. John Wiley & Sons, Inc.

Sweets, J., Dawes, R., & Monahan, J. (2000). Better decision through science. *Scientific American*, 82–87.

Tunbridge, W. M. et al. (1997). The spectrum of thyroid disease in a community: the Eidkham survey. *Clinical Endocrinology, 7*, 481–493.

Wong, E. T., & Steffes, M. W. (1984). A fundamental approach to the diagnosis of diseases of the thyroid gland. *Clinical Laboratory Medicine, 4*, 655–670.

Zhang, G., & Berardi, V. (1998). An investigation of neural networks in thyroid function diagnosis. *Health Care Management Science, 1*, 29–37.