International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:3, 2009

Unsupervised Feature Selection Using Feature Density Functions

Mina Alibeigi, Sattar Hashemi, Ali Hamzeh

***Abstract*—**Since dealing with high dimensional data is computationally complex and sometimes even intractable, recently several feature reductions methods have been developed to reduce the dimensionality of the data in order to simplify the calculation analysis in various applications such as text categorization, signal processing, image retrieval, gene expressions and etc. Among feature reduction techniques, feature selection is one the most popular methods due to the preservation of the original features.

In this paper, we propose a new unsupervised feature selection method which will remove redundant features from the original feature space by the use of probability density functions of various features. To show the effectiveness of the proposed method, popular feature selection methods have been implemented and compared. Experimental results on the several datasets derived from UCI repository database, illustrate the effectiveness of our proposed methods in comparison with the other compared methods in terms of both classification accuracy and the number of selected features.

***Keywords*—**Feature, Feature Selection, Filter, Probability Density Function

I. Introduction

SINCE data mining is capable of finding new useful information from datasets, it has been widely applied in various domains such as pattern recognition, decision support, signal processing, financial forecasts and etc [1]. However by the appearance of the internet, datasets are getting larger and larger which may lead to traditional data mining and machine learning algorithms to do slowly and not efficiently. One of the key solutions to solve this problem is to reduce the amount of data by sampling methods [2], [3]. But in many applications, the number of instances in the dataset is not too large, whereas the number of features in these datasets is more than one thousands or even more. In this case, sampling is not a good choice. Theoretically, having more features, the discrimination power will be higher in classification. However, this theory is not always true in reality since some features may be unimportant to predict the class labels or even be irrelevant [4], [5]. Since many factors such as the quality of the data, are responsible in the success of a learning algorithm, in order to extract information more efficiently, the dataset should not contains irrelevant, noisy or redundant features [6]. Furthermore, high dimensionality of data may cause the "curse of dimensionality" problem [7]. Feature reduction (dimensionality reduction) methods are one of the key solutions to all these problems.

Feature reduction refers to the problem of reducing the dimension by which the data is described [8]. The general purpose of these methods is to represent data with fewer features to reduce the computational complexity whereas preserving or even improving the discriminative capability [8]. Since feature reduction can brings a lot of advantages to learning algorithms, such as avoiding over-fitting and robustness in the presence of noise as well as higher accuracy, it has attracted a lot of attention in the three last decades. Therefore, vast variety of feature reduction methods was suggested which are totally divided into two major categories including feature extraction and feature subset selection. Feature extraction techniques projects data into a new reduced subspace in which the initial meaning of the features are not kept any more. Some of the well-known state-of-the-art feature extraction methods are principal component analysis (PCA) [5], non-linear PCA [12] and linear discriminant analysis (LDA) [12]. In comparison, feature selection methods preserve the primary information and meaning of features in the selected subset. The purpose of these schemas is to remove noisy and redundant features from the original feature subspace [12]. Therefore, due to preserving the initial meaning of features, feature selection approaches are in more of interest [8], [9].

Feature selection methods can be broadly divided into two categories: filter and wrapper approaches [9]. Filter approaches choose features from the original feature space according to pre-specified evaluation criterions, which are independent of specified learning algorithms.

Conversely, wrapper approaches selects features with higher prediction performances estimated according to specified learning algorithms. Thus wrappers can achieve better performance than filters. However, wrapper approaches are less common than filter ones because they need higher computational resources and are often intractable for large scale problems [9]. Due to its computational efficiency and independency to any specified learning algorithm, filter approaches are more popular and common for high

M. Alibeigi is with IT, Computer Science and Engineering Department, Shiraz University, Iran (corresponding author to provide phone: +98-711-6133544; fax: +98-711-6474605; e-mail: alibeigi@cse.shirazu.ac.ir).

S. Hashemi is with IT, Computer Science and Engineering Department, Shiraz University, Iran (e-mail: s_hashemi@shirazu.ac.ir).

A. Hamzeh is with IT, Computer Science and Engineering Department, Shiraz University, Iran (e-mail: hamzeh@shirazu.ac.ir).

dimensional datasets [9].

In this study, we present a new filter unsupervised feature selection algorithm which has the benefits of filter approaches. The proposed approach chooses more informative features according to their probability density function (pdf) relations. The main idea of the proposed scheme is firstly approximating the pdf of each feature independently in an unsupervised manner and then removing those features which their pdfs have higher covering areas with the pdfs of other features which are known as redundant features.

The rest of this paper is organized as follow. Section 2 discusses the related researches for unsupervised feature selection. Section 3 explains the proposed method for unsupervised feature selection applications. Our experimental results are given in section 4 and section 5 concludes the paper by a conclusion part.

II. Related Work

Conventional feature selection methods evaluate various subsets of features and select the best subset among all with the best evaluation according to an effective criterion related to the application. These methods often suffer from high computational complexity through their searching process when applied to large datasets. The complexity of an exhaustive search is exponential in terms of the number of features of the dataset. To overcome these shortcomings, several heuristic schemas have been proposed such as Branch and Bound (B&B) method which guarantees to find the optimal subset of features with computational time expectedly less than the exponential under the monotonicity assumption [11]. B&B starts from the full set of features and removes features by a depth first search strategy until the removing of one feature can improve the evaluation of the remaining subset of features [11]. Another popular approach is Sequential Forward Selection (SFS) which searches to find the best subset of features in an iterative manner starting from the empty set of features. In each step, SFS adds the feature to the current subset of selected features which yields to maximize the evaluation criterion for the new selected feature subset [12]. However, heuristic approaches are simple and fast with /quadratic complexity, but they often suffer from lack of backtracking and thus act poorly for nonmonotonic criterions. In [23], another heuristic method called Sequential Floating Forward Selection (SFFS) was proposed which performs sequential forward selection with the backtracking capability at the cost of higher computational complexity.

The former methods can be applied in both supervised and unsupervised schemas according to their evaluation criteria. Since the interest of this paper is developing an unsupervised feature selection method, here, we investigate only the unsupervised methods. These methods can be generally divided into two divisions: filter and wrapper approaches [4], [8], [12]. The principle of wrapper approaches is to select subset of features regarding a specified clustering algorithm. These methods find a subset of features on which when the specified clustering algorithm was trained, it achieve the highest performance according to a pre-specified criterion. Some examples of these approaches are [13]-[16]. Conversely, filter methods select features according to an evaluation criterion independent of specified clustering algorithm. The goal of these methods is to find irrelevant and redundant features and remove them from the original feature space. In order to find irrelevant and redundant features, various dependency measures have been suggested such as correlation coefficient [6], linear dependency [17] and consistency measures [18].

In this paper, we propose a feature subset selection based on the probability density function of features which is able to handle the nonlinearity dependency between features in an unsupervised framework. The following section explains the proposed method in details.

III. The Proposed Method for Unsupervised Feature Selection

The proposed unsupervised feature selection which is a filter approach includes four steps. Fig. 1 illustrates the whole process of the proposed feature selection method. The proposed method finds the relation between each two features as if they are similar or not according to their estimated probability density functions and removes those features which are more similar to other features as redundant features because all or most of their information is repeated in other features.

As was shown in Fig. 1, the first step in our feature selection approach is estimating the probability density function of each feature. The methods for estimating probability density functions can be totally categorized into parametric and non-parametric approaches [20]. The parametric methods assume a particular form for the density, such as Gaussian, so that only the parameters (mean and variance) need to be determined. In comparison, non-parametric methods do not assume any knowledge about the density of the data and computes the density directly from the instances and because of this reason they are in more of interest. The general form of non-parametric estimation of probability density functions is according to the following formula:

$$p(x) \cong \frac{k}{N * V} \tag{1}$$

where, $p(x)$ is the value of estimated probability density function for instance x, V is the volume surrounding x, N is the total number of examples or instances and k is the number of examples inside V. Two basic approaches can be adapted to practical non-parametric density estimation methods based on the status of k and V. Fixing the value of k and determining the corresponding volume V from the data, leads to methods commonly referred to as *K Nearest Neighbor (KNN)* methods. On the other hand, when the value of the volume V is chosen to be fixed and k is determined from the data, the non-

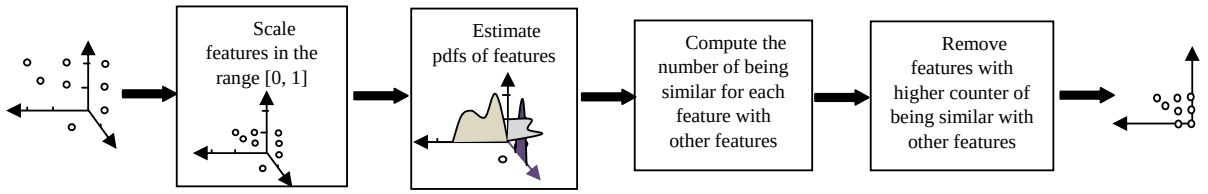

Fig. 1 The whole process of the proposed feature selection method.

parametric estimation method is called *Kernel Density Estimation (KDE)*. Generally, the estimates that can be obtained with the KNN approaches are not very satisfactory because of some drawbacks. The estimates by KNN methods are prone to local noise with very heavy tails. Moreover, the resulting density is not a true probability density since its integral over all the sample space diverges [24]. In spite of these reasons, in this study, we estimate probability density functions through the KDE method with Gaussian kernel. It is noted that our proposed feature selection algorithm is not sensitive to any particular estimation method. However, using more accurate estimation methods cause the algorithm to perform more efficiently.

In order to compare pdfs of different features, all feature values are scaled into the [0, 1] interval because the range of various features may be different. Afterwards, the probability density functions for each of the features are computed according to KDE methods.

Having estimated the probability density functions for each feature, the similarity between each of the two features is calculated. Two features are considered as similar features if the Mean Square Error (MSE) of their pdfs is less than a user specified threshold. Similar features contain nearly the same information because their pdfs are sufficiently similar. Thus, one of the similar features can be removed without a considerable loss of information. Among similar features, features which are similar to more other feature of the whole feature space are removed. Removing features with higher frequencies of being similar with other features, there is a higher probability that the feature which is selected to be removed, contributes its information with other features. Therefore, by removing this feature the loss of information will be at least. Algorithm 1 represents the steps of the proposed feature selection approach.

IV. Experimental Results and Discussion

The comparisons were carried out in three datasets coming from the UCI Machine Learning Repository including Ecoli, Ionosphere and Sonar. Table I shows a summary of the characteristics of these datasets used in this paper to assess the performance of the proposed method.

In order to evaluate the performance of a feature selection method, the accuracy of the classifiers trained on those features selected by the mentioned feature selection method must be compared with the accuracy of classifiers trained on the full set of features named as All Features. There are many classifiers in machine learning domains with different biases. The most well-known classifiers for evaluation of the feature selection methods are *Naive Bayes* (*NB*) [10] and *K Nearest Neighbor* (*KNN*) classifiers [12]. Naïve Bayes is a simple probabilistic classifier based on the assumption of class conditional independence of features. K Nearest Neighbor is a lazy learning algorithm which classifies each new test example based on its K nearest training examples. In this paper, we evaluate the performance of different feature selection methods based on their accuracies on these classifiers.

TABLE I
DATASETS USED IN THIS PAPER FOR EVALUATIONS.

Dataset	*#Instance*	*# Features*	*# Class*
Sonar	*208*	*60*	*2*
Ionosphere	*351*	*34*	*2*
Ecoli	*336*	*7*	*8*

To show the effectiveness of our proposed method, we compared our method with supervised approaches proposed by Hall *et al.* [19] and Lie *et al.* [4] named as Correlation-based Feature subset Evaluation (CfsSubsetEval) and Consistency-based feature Subset Evaluation (ConsistencySubsetEval), respectively. We also compared our method with an unsupervised Sequential Forward Selection (SFS) scheme for which Entropy is used as the evaluation criterion. The entropy is defined as follows:

$$Entropy = -\sum_{p=1}^{l}\sum_{q=1}^{l}(sim(p,q)*\log(sim(p,q)) + (1-sim(p,q))*\log(1-sim(p,q)))$$

$$Sim(p,q) = e^{-\alpha D_{pq}} \qquad (2)$$

$$D_{pq} = [\sum_{j=1}^{M}(\frac{x_{p,j} - x_{q,j}}{\max_j - \min_j})^2]^{1/2}$$

where D_{pq} is distance between two points of p and q and $x_{p,j}$ denotes feature value for p along with the jth feature. max_j and min_j are the maximum and minimum values overall value along jth feature and N denotes the number of features. In (2), α is a positive constant which is set as $\alpha = \frac{-\ln 0.5}{\overline{D}}$ where $\overline{D}$ is the average distance between all data points.

ALGORITHM I
THE STEPS OF THE PROPOSED UNSUPERVISED FEATURE SELECTION METHOD.

Unsupervised Feature Selection using statistical measurements

Input: *$D=\{d_1, d_2, ..., d_N\}$ // a data set containing N items*
Input: *$F=\{f_1, f_2, ..., f_n\}$ // a data set has N features*
Output: *$F^{(S)}$ // the feature subset identified by proposed method*

Begin
Scale each feature in the range [0,1]
Estimate the probability density function of each feature
For i=1:num_features-1
For j=i+1:num_features
Calculate MES(density of feature i, density of feature j)
If MSE <= epsilon
Consider features i and j to be similar
Increment the similarity of features i and j's corresponding counter
Remove those features with the higher counter from the provided list of similar features,
Return the rest of features left in the feature list
End

Tables 2-4, illustrate the experimental results on the introduced datasets separately. As the results show in Tables and schematically in Figs. 2-5, the performance of the proposed method is similar to All Features or even better due to removing redundant features. By considering this point into account that in average, the proposed method selects only a half of the features while All Features method uses all features (see Fig. 6). Also, our method is comparable to CfsSubsetEval in terms of the average accuracy on different classifiers while our method selects more features. However, it is noticeable that our method is an unsupervised method which has access to less information in comparison with the CfsSubsetEval which is a supervised approach and has access to the class labels. Furthermore, our method has higher accuracy in comparison to the supervised ConsistencySubsetEval and unsupervised SF with Entropy methods. In general, it can be concluded that the proposed method is more efficient than the unsupervised SF with Entropy method and comparable to the supervised schemas.

TABLE II
EXPERIMENTAL RESULTS ON IONOSPHERE DATASET BY THE PROPOSED AND COMPARED METHODS IN TERMS OF NUMBER OF SELECTED FEATURES AND CLASSIFICATION ACCURACY.

Feature Selection Method	*# Selected Features*	*NB (Accuracy)*	*IB1 (Accuracy)*
All Features	*34*	*82.6211*	*86.3248*
CfsSubsetEval	*14*	*92.0228*	*88.8889*
ConsistencySubsetEval	*7*	*87.1795*	*87.7493*
SFS with Entropy	*14*	*78.3476*	*80.3419*
Proposed Method	*12*	*92.0228*	*91.1681*

TABLE III
EXPERIMENTAL RESULTS ON ECOLI DATASET BY THE PROPOSED AND COMPARED METHODS IN TERMS OF NUMBER OF SELECTED FEATURES AND CLASSIFICATION ACCURACY.

Feature Selection Method	*# Selected Features*	*NB (Accuracy)*	*IB1 (Accuracy)*
All Features	*7*	*85.4167*	*80.3571*
CfsSubsetEval	*6*	*85.4167*	*80.0595*
ConsistencySubsetEval	*6*	*85.4167*	*80.0595*
SFS with Entropy	*6*	*79.4643*	*76.7857*
Proposed Method	*6*	*85.4167*	*80.0595*

TABLE IV
EXPERIMENTAL RESULTS ON SONAR DATASET BY THE PROPOSED AND COMPARED METHODS IN TERMS OF NUMBER OF SELECTED FEATURES AND CLASSIFICATION ACCURACY.

Feature Selection Method	*# Selected Features*	*NB (Accuracy)*	*IB1 (Accuracy)*
All Features	*60*	*67.7885*	*86.5385*
CfsSubsetEval	*19*	*67.7885*	*83.6538*
ConsistencySubsetEval	*14*	*66.8269*	*85.0962*
SFS with Entropy	*19*	*58.6538*	*66.8269*
Proposed Method	*16*	*68.75*	*85.5769*

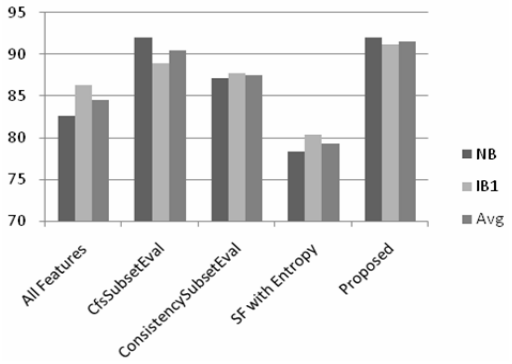

Fig. 2 Results on Ionosphere dataset.

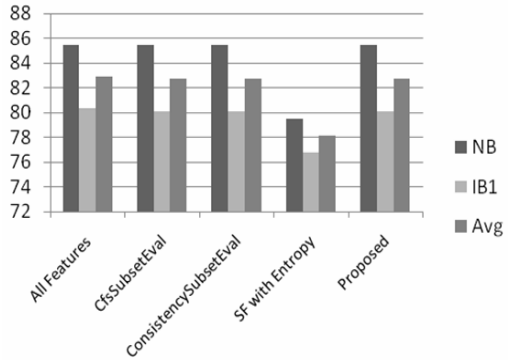

Fig. 3 Results on Ecoli dataset.

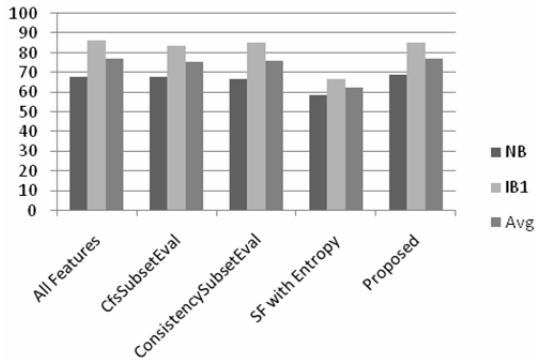

Fig. 4 Results on Sonar dataset.

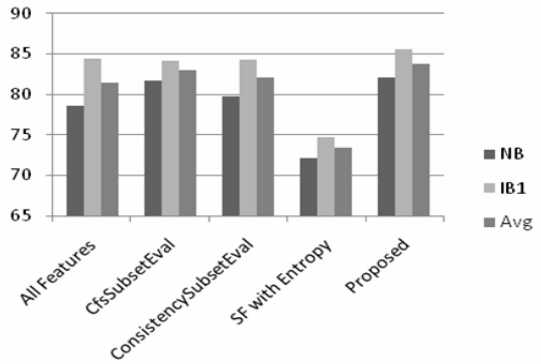

Fig. 5 Results on All datasets averaged.

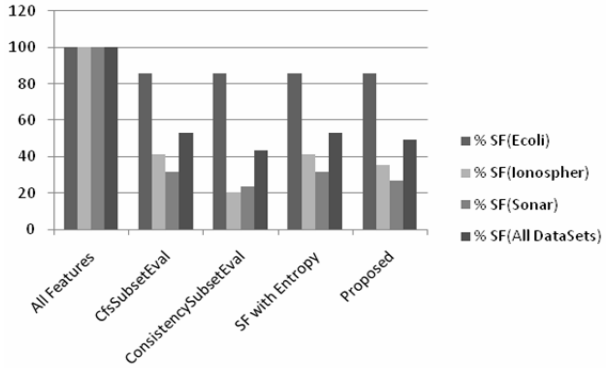

Fig. 6 The percentage of selected features for each dataset and in average for all datasets for the proposed and compared methods.

V. Conclusion

Feature selection techniques have a key role when encountering high dimensional datasets. Recently, filter based feature selection methods are of more interest because of their independence to any particular learning algorithm and their fastness. Therefore, in this study, we proposed a new filter unsupervised feature selection scheme which selects features based on the estimation of their probability density functions and the relation between each feature pdf with other feature pdfs. The main idea is that a feature which is similar to most of the features is redundant because all or most of its information is repeated in those similar features. So, this feature can be removed from the original feature space with the least loss of information. Experimental results show that the proposed method can find the subset of features with more informative features in comparison with the unsupervised feature selection method. Also, its results are comparable to the supervised feature selection frameworks.

For future work, it might be useful to apply this idea in the field of supervised feature selection methods and find the probability density function of each class in each feature separately and finds the similarity between features by considering their densities over different classes.

Acknowledgment

This work was supported by the Iran Tele Communication Research Center.

References

[1] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, “From data mining to knowledge discovery in databases”, AI Magazine, vol. 17, 1996, pp. 37–54.
[2] M. Lindenbaum, S. Markovitch, D. Rusakov, “Selective sampling for nearest neighbor classifiers”, Machine learning, vol. 54, 2004, pp. 125–152.
[3] A.I. Schein, L.H. Ungar, “Active learning for logistic regression: an evaluation”, Machine Learning, vol. 68, 2007, pp. 235–265.
[4] M.A. Hall, “Correlation-based feature subset selection for machine learning”, Ph.D. Dissertation, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1999.
[5] I.K. Fodor, “A survey of dimension reduction techniques”, Technical Report UCRL- ID-148494, Lawrence Livermore National Laboratory, US Department of Energy, 2002.
[6] M.A. Hall, “Correlation-based feature selection for discrete and numeric class machine learning”, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 2000.
[7] R. Bellman, “Adaptive Control Processes: A Guided Tour”, Princeton University Press, Princeton, 1961.
[8] H. Liu, J. Sun, L. Liu H. Zhang, “Feature selection with dynamic mutual information”, Pattern Recognition, vol. 42, 2009, pp. 1330 – 1339.
[9] N. Pradhananga, “Effective Linear-Time Feature Selection”, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 2007.
[10] George H. John, Pat Langley: Estimating Continuous Distributions in Bayesian Classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 1995, pp. 338-345.
[11] M.P. Narendra, K. Fukunaga,“A branch and bound algorithm for feature subset selection”, IEEE Trans. Comput. Vol. 26, 1997, pp. 917–922.
[12] P.A. Devijver, J. Kittler, “Pattern Recognition: A Statistical Approach”, Englewood Cliffs: Prentice Hall, 1982.
[13] M. Dash, H. Liu, “Unsupervised Feature Selection”, Proc. Pacific Asia conf. Knowledge Discovery and Data Mining, 2000, pp. 110-121.
[14] J. Dy, C. Btodley, “Feature Subset Selection and Order Identification for Unsupervised Learning”, Proc. 17th Int’l. Conf. Machine Learning, 2000.
[15] S.Basu, C.A. Micchelli, P. Olsen, “Maximum Entropy and Maximum Likelihood Criteria for Feature Selection from Multi-variate Data”, Proc. IEEE Int’l. Symp. Circuits and Systems, 2000, pp. 267-270.
[16] S.K .Pal, R.K. De, J. Basak, “Unsupervised Feature Evaluation: A Neuro-Fuzzy Approach”, IEEE Trans. Neural Network, vol. 11, 2000, pp. 366-376.
[17] S.K .Das, “Feature Selection with a Linear Dependence Measure”, IEEE Trans. Computers, 1971, pp. 1106-1109.
[18] G.T. Toussaint, T.R. Vilmansen, “Comments on Feature Selection with a Linear Dependence Measure”, IEEE Trans. Computers, 1972, 408.

[19] H. Liu, R. Setiono: “A probabilistic approach to feature selection - A filter solution”. In: 13th International Conference on Machine Learning, 1996, pp. 319-327.
[20] K. Fukunaga, “Introduction to Statistical Pattern Recognition”, Academic Press, 2nd Ed. 1990.
[21] E. Frank, M.A. Hall, G. Holmes, R. Kirkby, B. Pfahringer, “Weka - a machine learning workbench for data mining”, In The Data Mining and Knowledge Discovery Handbook, Springer 2005, pp. 1305-1314.
[22] M. Dash, H. Liu, “Unsupervised Feature Selection”, Proc. Pacific Asia Conf. Knowledge Discovery and Data Mining, 2000, pp. 110-121.
[23] P. Pudil, J. Novovicova,J. Kittler, “Floating Search Methods in Feature Selection”, Pattern Recognition Letters, vol. 15, 1994, pp. 1119-1125.
[24] R.O. Duda, P.E. Hart, D.G. Stork, “Pattern Classification”, Second Edition, Wiley, 1997.

Mina Alibeigi was born in Shiraz, Iran in 1986. She received her B.Sc. degree in Computer Engineering from Shiraz University in 2008. She is currently an M.Sc. student in Artificial Intelligence at Shiraz University.

Her research interests include dimension reduction, learning, clustering, bio-inspired algorithms and neuroscience.

Sattar Hashemi received the PhD degree in Computer Engineering from the Iran University of Science and Technology, in conjunction with Monash University, Australia, in 2008. He is currently a lecturer in the Electrical and Computer Engineering School, Shiraz University, Shiraz, Iran. His research interests include data stream mining, database intrusion detection, dimension reduction, and adversarial learning.

Ali Hamzeh received his Ph.D. in artificial intelligence from Iran University of Science and Technology (IUST) in 2007. Since then, he has been working as assistant professor in CSE and IT Department of Shiraz University. There, he is one of the founders of local CERT center which serves as the security and protection service provider in its local area. As one of his research interests, he recently focuses on cryptography and steganography area and works as a team leader in CERT center to develop and break steganography method, especially in image spatial domain. Also, he works as one of team leaders of Soft Computing group of shiraz university working on bio-inspired optimization algorithms. He is co-author of several articles in security and optimization.