Multifactor Affiliation Analysis: A Multifactor Dimensionality Reduction based Learning Model for Knowledge Discovery and Similarity Measure in 2-way Data Classification

Aditya C.R.
Department of Computer Science and Engineering,
VVIET, Mysuru-570028,
Karnataka, India

M.B. Sanjay Pande
Department of Computer Science and Engineering,
GMIT, Davangere-577004,
Karnataka, India

ABSTRACT

Extracting useful information from the datasets of high dimension and representing the learnt knowledge in an efficient way is a challenge in knowledge discovery and data mining. Although many pattern recognition, knowledge discovery and data mining techniques are available in literature, there is a need for techniques that represent the high dimensional data in a low dimension by preserving useful information for supervised learning. In this work, we design a novel model which effectively captures both inter-feature and intrafeature relationships in the sample space for knowledge discovery by performing dimensionality reduction, using a modified version of multi-factor dimensionality reduction based algorithm. The model uses the learnt knowledge to quantify the similarity of a test sample with respect to a specific class. The evaluation of the model on Fisher's IRIS dataset containing 50 samples each from three types of IRIS species- setosa, versicolor and verginica, shows that the designed model explores the data set for useful information and assigns test samples to a specific class with measurable similarity indices.

Keywords

Multifactor Dimensionality Reduction, Knowledge Discovery, Similarity Measure, Classification

1. INTRODUCTION

Knowledge discovery is a nontrivial extraction of implicit class present in unknown and potentially useful information from data. Knowledge discovery and data mining are closely related to fields such as machine learning, statistics, and databases [6], [7]. The growth in the size and number of existing databases far exceeds human abilities to analyze such data, thus creating both a need and an opportunity for extracting knowledge from databases. knowledge discovery and data mining has been ranked as one of the most promising research topics for both databases and machine learning researchers [4] [8], [9]. Advances in data collection techniques have generated a need for new techniques and tools that can intelligently and automatically assist in transforming the data into useful knowledge [8], [13]. A database usually stores a large amount of data, of which only a portion may be relevant to specific learning task. For example to characterize the features of graduate students in science, only the data relevant to graduates in science are appropriate for the learning process [11], [16]. Extensive amounts of information stored in databases have created the need for development of specialized tools for storing, accessing, analysis, and effectiveness usage of stored knowledge and data. Efficient algorithms and often crude approximations play an important role in analysing high dimensional and complex data [10].

Traditional parametric- statistical methods are less practical in dealing with high dimensional data. High order interactions between the features are handled poorly by parametric methods and often create many contingency table cells that contain no observations. This can lead to very large coefficient estimates and standard errors. Further, modeling high order interactions using standard statistical methods can be cumbersome and often leads to difficulties in model building and interpretation of parameters. The limitations of traditional statistical modeling techniques have led to development and application of novel statistical methods for handling high dimensional data and to uncover complex relationships between the data features. Multifactor dimensionality reduction (MDR), combinatorial partitioning method are some of such prominent methods [1], [15] ,[14], [18]. These methods are model free and non-parametric in nature, designed to uncover complex relationships without relying on any specific model. Multifactor dimensionality reduction method classifies the normal and abnormal samples in an ad hoc manner based on a simple comparison of standard-specific ratios from a balanced standard-specific training dataset. Development of good subjective or objective measures of interestingness of discovered patterns is one of the central problems in the field of knowledge discovery [19]. One of the most common representations of patterns and measuring similarity is through binary feature vectors, which can be used to design distance measures that play a critical role in many clustering and classification problems [3], [5]. There are many machine learning techniques available to predict group membership for given data instances [2], [17].

The proposed model uses the standard-specific ratio comparison concept of MDR [20] to reduce the m * n feature space into m * 1 dimensional space. A *knowledge base*, which captures the complex inter-feature relationships and the feature interaction weight of the sample, is generated as the feature dimension is reduced. The classification of a test sample is carried out by comparing the feature values of the test sample with the knowledge base. Similarity index of the test sample to the specific reference base is measured by using a Euclidean distance based novel similarity measure.

The proposed model consists of two phases :

(1) ***Knowledge-base creation:*** This initial phase is a learning phase, where the complex relationships between the sample features are extracted from the training dataset using an MDR based algorithm. The extracted features which represent the learnt knowledge about the data are stored in a table which forms the knowledge base.
(2) ***Affiliation analysis:*** : The next phase affiliates the closeness of the test sample to a specific class. The test sample feature values are compared with the knowledge base and the affiliation is carried out by using a novel distance measure.

2. MATERIALS AND METHODS

2.1 Data-Set

Proposed methodology is illustrated using Fisher's IRIS data set [12], which is a classical data from measurement of three species of Iris i.e., Iris setosa, Iris Verginica, Iris Versicolor. With 50 samples in each group the four features measured are petal width, petal length, sepal width and sepal length. Iris setosa is generally considered to be standard reference base and, Iris Verginica, Iris Versicolor as specific data sets of various degrees.

2.2 Knowledge-base creation

The proposed model extracts complex relationships between the features and creates a knowledge base in the form of "Feature Interaction Table" (FIT). FIT effectively captures the complex relationships between the features of a dataset in the form of a value-weight pair. Initially the feature values are converted to a value 0 or 1, depending on whether the values are close to the mean of standard or specific samples respectively. For any arbitrarily chosen y-features from a given sample, there exists 2^y possible combinations of values in the converted feature space. For example, (0, 0) (0, 1) (1, 0) (1, 1) are the four possible combinations between two selected features' values. The ratio of occurance for each of these features' value combinations in standard and specific samples is compared, and selected features are merged to form a new feature. The *value* of new merged feature (*feature interaction value*) for each combination of old features' value combinations along with the *feature interaction weight* calculated using equation-1 is updated in the FIT. This process is repeated till m x n dimensional dataset gets reduced to m x 1 dimensional dataset. FIT thus created for the above said converted feature space will contain a row for each possible features' value combinations and a column for every new feature created. So a FIT generated for an m x n dataset will have 2^y rows and $n - 1$ columns. The interpretation of FIT is summarized in table-1.

Feature interaction weight (FIW) updated for each possible features' value combinations quantifies the degree of occurance of those value combinations with respect to specific samples as compared with standard samples. The weights are computed based on the observation of a subset with relavence to the features in concern. FIW is designed such that it varies between zero to one, where one is the highest quantifying factor for a value combination with respect to specific samples. let $specific_{occurance}$ represent the occurance count of a features' value cambination in standard samples and $standard_{occurance}$ represent the same in standard samples. Then FIW must exhibit the following properties:

property 1: If $specific_{occurance} = 0$ and $standard_{occurance} = m$, then $FIW = 0$. A value zero for FIW indicates that the occurance of a features' value combination is unique to standard samples and the specific samples do not exhibit that pattern of combination.

property 2: If $specific_{occurance} = m$ and $standard_{occurance} = 0$, then $FIW = 1$. If the degree of occurance of a features' value combination is unique and at its maximum value for specific samples, then FIW must be equal to one.

property 3: If $specific_{occurance} >= standard_{occurance}$ then $0.5 < FIW < 1$, else $0 < FIW < 0.5$. As the occurance of a feature combination pattern increases in specific samples then the value of FIW increases towards 1 else it decreases towards 0.

The training data set used in knowledge-base creation consists of 2*m samples with "m" equal standard and specific samples. So the factors $specific_{occurance}/(2 * m)$ and $standard_{occurance}/(2 * m)$ can never be greater than 0.5. Further the FIW should quantify the occurance of the value combinations with respect to specific samples as compared with standard samples, and must be equal 1 if such an observed combination is strongly affiliated to the specific class. Since only 50% of the taining samples are specific samples, a value 0.5 is added as a term to the equation-1 and 0.5 weightage is given to the factor $standard_{occurance}/(2 * m)$.

$$FIW = 0.5 + [\frac{Specific_{occurance}}{2 * m} - 0.5 * \frac{Standard_{occurance}}{2 * m}] \tag{1}$$

2.2.1 *FIT generation algorithm.*

The steps for creating the feature-interaction table by performing dimensionality reduction are as follows.

Input: n-dimensional feature values [f1 , f2 , . . . fi ,. . . fn,] for m specific and m standard samples.
Output: Feature Interaction Table with (interaction-value, interaction-weight) pair.

Step 1: Calculate μ_{ni} and μ_{ai} for each feature f_i in the sample space, where μ_{ni} and μ_{ai} denote the ith feature value mean of standard and specific samples respectively.

Step 2: Convert all the training sample feature values to either 0 if they are close to the value of μ_{ni} or to 1 if the values are nearer to μ_{ni}.

Step 3: For any selected two features in the new converted sample space, calculate the occurance count of each possible features' value combinations.

Step 4: Using the occurance count calculated in step-2, merge the two features into one by comparing the standard-specific ratio of feature values. The value of new merged feature will be 1 or 0, if the combination of old feature values is predominant in specific samples or standard samples respectively.

Step 5: Calculate feature-interaction weight for all the four old features' value combinations using equation-1.

Step 6: Create a new column in FIT and update it with (interaction-value, interaction-weight) pair for all the four possible old features' value combinations.

Table 1. Interpretation of the values in feature-interaction table(FIT). FIT data Dj_i indicates data from row-j and column-i

FIT data	Feature interaction value	Affiliation of test/training sample		
		Inferred from previously observed features' values	Inferred from present feature value	Net/overall affiliation
D1i	0	standard	standard	standard
D2i	0	standard	specific	standard
D3i	0	specific	standard	standard
D4i	0	specific	specific	standard
D1i	1	standard	standard	specific
D2i	1	standard	specific	specific
D3i	1	specific	standard	specific
D4i	1	specific	specific	specific

Repeat steps 3 to 6 till the n-dimensional feature space is reduced to a single dimensional space.

2.3 Affiliation Analysis

Association/Dissociation of a sample with respect to degree of affiliation is done through calculated weights. The class assigned to the sample based on the previously observed i-1 features and the class assigned to the sample based on newly observed feature "i" will form the row index to the FIT for affiliation analysis of a test sample. The column index will be one less than the number of feature values assessed for affiliation analysis of the test sample. A value 0/1 obtained from the FIT for the given row and column indices will affiliate the test sample as standard/specific based on the previously assigned class and newly observed feature value. The weight obtained along with the FIT value quantifies the degree of similarity of a test sample with reference to the specific samples.

The affiliation algorithm used in our model estimates the degree of test sample being away from the standard samples on a measurable scale. When a test sample is presented for affiliation, the feature values are initially compared with the respective feature means caluculated during knowledge base creation stage for training samples, and are converted to a value 0 or 1 based on the closeness of the value to standard mean or specific mean respectively. This convertion step will affiliate the sample to either standard or specific class for each feature. But the overall affiliation is employed by using a novel distance measure: *Specific-Affiliation distance* which is derived from a variant of eucledian distance measure.

The Eucledian distance [21] between two points (x,y) and (p,q) is given by

$$Eucledian_{distance} = \sum \sqrt{(x-p)^2 + (y-q)^2} \qquad (2)$$

We use the idea behind eucledian distance to find the feature wise distance of a sample with respect to the mean of the feature in consideration. Let μ_i denote the mean value for the i^{th} feature in a data set containing equal number of standard and specific samples. Given the mean value μ_i along with the feature value f_i sample , the *Feature Distance* is calculated using equation-3. The term $(\mu_i - f_i)$ denotes the farness/closeness of a sample for a selected feature with respect to the mean value. The *Net Feature Distance* for a sample is the sum of all the individual feature distances, or in other words it is the final *Feature Distance* after processing all the feature values ($i = n - 1$ in equation-3).

$$FD_{i+1} = \sqrt{(FD_i)^2 + (\mu_{i+1} - f_{i+1})^2} \qquad (3)$$

The *Specific-Affiliation distance* is given by

$$SAD = \frac{Affiliation_{weight}}{n-1} * Feature_{distance}(net) \qquad (4)$$

The $Affiliation_{weight}$ gives overall weightage for the degree of affiliation of test sample to the specific class. The features of the test samples are merged iteratively into one new feature by comparing the values with the FIT, and the *Feature interaction weight* for the feature values merged in each iteration is added to get the $Affiliation_{weight}$.

The overall steps involved in affiliation analysis are as follows.
Input: Feature Interaction Table and n-dimensional feature values [f1 , f2 , . . . fi ,. . . fn,] for a test sample.
Output: Specific-Affiliation distance for the test sample.

Step 1: calculate the net feature distance for the test sample by iteratively summing up the $Feature_{distance}$ for each feature calculated using equation-3

Step 2: Convert all the test sample feature values $(f_1, f_2 ... f_i)$ to either 0 or 1, based on the closeness of value to the mean of standard μ_{ni} or specific μ_{ai} samples calculated in FIT generation phase. Set $Affiliation_{weight}$ to 0.

Step 3: Select two features f_i and f_{i+1}, and access the FIT based on the selected two feature values: *(row index)* and number of features merged: *(column index)*.

Step 4: Merge the two selected features into a new feature. New merged value will be the *Feature interaction value* present in the accessed cell of FIT.

Step 5: Add the *Feature-interaction weight* in the FIT entry to $Affiliation_{weight}$. Repeat steps 2 to 5 till the n-dimensional feature space of test sample is reduced to a single dimensional space.

Step 6: Calculate the *Specific-Affiliation distance* by using equation-4

3. RESULTS

The designed model is evaluated on "IRIS" data. Two sets of experimentation was carried out. In the first experimentation iris-setosa was considered as standard sample and iris-versicolor as specific. 50 samples of iris-setosa and 50 samples of iris-versicolor were

Table 2. Feature Interaction table for 50 setosa and 50 versicolor samples

0 , 0.46	0 , 0.01	0 , 0.02
1 , 0.50	1 , 0.91	0 , 0.48
0 , 0.46	0 , 0.49	1 , 0.50
1 , 0.58	1 , 0.59	1 , 1.00

Table 3. Feature interaction table for 50 versicolor and 50 verginica samples

1 , 0.50	0 , 0.45	1 , 0.50
0 , 0.45	1 , 0.50	0 , 0.39
1 , 0.50	0 , 0.44	1 , 0.50
1 , 0.55	1 , 0.61	1 , 0.61

Table 4. Range of specific affiliation distance obtained for 50 test samples containing 25 samples of both setosa and versicolor. Setosa is considered as standard sample and versicolor as specific

	Specific affiliation distance	
sample	min	max
setosa	18.579836	28.621592
versicolor	120.126701	207.886948

chosen as standard-specific samples for generating the FIT. For the affiliation analysis, 25 samples of setosa and 25 samples of versicolor were chosen randomly. Table-2 shows the FIT generated for the chosen training set and the Table-4 summarizes the specific affiliation distances obtained for test samples which comprised of setosa-versicolor smaples. Second experimentation was carried out by considering iris-verginica as standard sample and iris-versicolor as specific. FIT was generated by considering 50 samples of iris-verginica and 50 samples of iris-versicolor as standard-specific samples. For the affiliation analysis, 25 samples of verginica and 25 samples of versicolor were chosen randomly. Table-3 shows the FIT generated for the chosen training set and the table-5 summarizes the affiliation distances obtained for test samples which comprised of verginica-versicolor smaples.

Our model represents the feature values either by 0 or 1 based on the deviation of a value from the mean of respective training samples' features. Also the distance measure used to affiliate the test samples captures the essence of mean parameter and this helps in assessing the affiliation of a sample across each dimension. The weightage factors used in similarity measure are computed by comparing the standard-specific ratio of features' value combinations and they denote the extent to which the features influence the class of a sample by taking the inter-feature relationships into account. Among the three classes in the dataset used, setosa is generally considered to be standard reference base. Verginica and versicolor are considered to be specific reference bases of various degree. Hence setosa should be clearly separable from the other two specific classes. In the data set considered versicolor and verginica have overlapping feature values and are not clearly distinguishable with respect to certain features. The results obtained are in par with these facts. Setosa and versicolor samples are assigned to non-overlapping classes through separable range of similarity indices. But verginica and versicolor samples have overlapping affiliation distances and are not clearly classified into separate classes, hence proving the correctness of the designed model.

Table 5. Range of specific affiliation distance obtained for 50 test samples containing 25 samples of both versicolor and verginica. Versicolor is considered as standard sample and verginica as specific

	Specific affiliation distance	
sample	min	max
versicolor	117.888527	166.555237
verginica	149.159088	227.644394

4. CONCLUSION

Dimensionality reduction algorithms such as multifactor dimensionality reduction can be used to understand feature properties in a dataset and generate a knowledge base from the learnt parameters. The knowledge base created using the proposed model captured the interaction details among features and also the impact of each feature in deciding the class of a sample. A novel Distance measure used in affiliation analysis quantified the test samples by a range of clearly separable distance values for non-overlapping classes. Overall quantification of the samples are done by combining the individual weights derived from feature wise interaction in knowledge base creation stage.

The future work will extend the proposed model into a more generic one, to handle ”n” classes of data for knowledge base creation. A distance metric to assign samples into various classes by quantifying their similarity to a specific class among ”n” distinguishable classes through measurable similarity indices can be designed with reference to the knowledge base created.

5. REFERENCES

[1] Marylyn D Ritchie Alison A Motsinger. Multifactor dimensionality reduction: An analysis strategy for modelling and detecting gene gene interactions in human genetics and pharmacogenomics studies. *Human Genomics*, 2(5)(5):318–328, march 2006.

[2] Phipps Arabie, Lawrence J Hubert, and Geert De Soete. *Clustering and classification*. World Scientific, 1996.

[3] Seung-Seok Choi, Sung-Hyuk Cha, and Charles C Tappert. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48, 2010.

[4] E R Davies. Machine vision: theory, algorithms, practicalities. *Signal Processing and Its Applications*, 1996.

[5] Michel Marie Deza and Elena Deza. *Encyclopedia of distances*. Springer, 2009.

[6] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[7] Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, et al. Knowledge discovery and data mining: Towards a unifying framework. In *KDD*, volume 96, pages 82–88, 1996.

[8] Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. Advances in knowledge discovery and data mining. 1996.

[9] Jiawei Han. Data mining techniques. In *ACM SIGMOD Record*, volume 25, page 545. ACM, 1996.

[10] Markus Hegland. Data mining techniques. *Acta Numerica 2001*, 10:313–355, 2001.

[11] E. Dallas Johnshon. *Applied multivariate methods for data analysis*. Kanas University, Duxbury Press, 1998.

[12] M. Lichman. UCI machine learning repository, 2013.

[13] Oded Maimon and Lior Rokach. *Data mining and knowledge discovery handbook*, volume 2. Springer, 2005.

[14] Alison A Motsinger Marylyn D Ritchie. Multi factor dimensionality reduction for detecting gene-gene and gene-environment interactions in pharmacogenomics studies. *Pharmogenomics*, 2005.

[15] Alison A Motsinger and Marylyn D Ritchie. Multifactor dimensionality reduction: an analysis strategy for modelling and detecting gene-gene interactions in human genetics and pharmacogenomics studies. *Human genomics*, 2(5):318, 2006.

[16] George Nagy. Twenty years of document image analysis in pami. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1):38–62, 2000.

[17] Thair Nu Phyu. Survey of classification techniques in data mining. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1, pages 18–20, 2009.

[18] Marylyn D Ritchie and Alison A Motsinger. Multifactor dimensionality reduction for detecting gene-gene and gene-environment interactions in pharmacogenomics studies. 2005.

[19] Avi Silberschatz and Alexander Tuzhilin. What makes patterns interesting in knowledge discovery systems. *Knowledge and Data Engineering, IEEE Transactions on*, 8(6):970–974, 1996.

[20] Wikipedia. Multifactor dimensionality reduction — wikipedia, the free encyclopedia, 2014. [Online; accessed 15-September-2015].

[21] Wikipedia. Euclidean distance — wikipedia, the free encyclopedia, 2015. [Online; accessed 15-September-2015].