6.7 SOM APPLICATION: SIMILARITY BETWEEN EUROPEAN LANGUAGES

Many practical SOM applications are concerned with clustering of multivariate data. In these applications, the goal is to model high-dimensional data in a two-dimensional feature space. Then the distance relationships in the feature space can be used to infer similarity between data samples. The two-dimensional structure of SOM enables visualization and human interpretation of high-dimensional data. Of course, SOM modeling methodology assumes that the distance metric in the input space is meaningful and informative. So the success of SOM clustering (or any other clustering method) depends to a large degree on the proper encoding and representation of high-dimensional data. This data encoding is application-dependent and requires common sense and good engineering.

There is a great diversity and variation among the world's human languages, and many languages appear, at first glance, completely unrelated. However, scientists and linguists have come to recognize that all human languages come from just a few prehistoric sources. For example, most European languages belong to a family of Indo-European languages. The similarity among several European languages and Sanskrit is evident from the fact that the word 'mother' starts with the same letter 'm' in all European languages. Comparative linguistics and historical linguistics are concerned with comparing languages in order to establish their similarity and historical relatedness. The comparison should use all linguistic aspects of languages, including phonology (language-specific patterns of sound), morphology (structure of words), syntax and lexicon. Based on such analysis, most European languages are known to belong to two language families, Indo-European and Finno-Ugric. Within the Indo-European family, there are several well-defined language groups, such as Germanic, Slavic and Romance groups.

This study investigates the possibility of establishing relations between European languages based on rough similarity measure between contemporary words. Further, this relative degree of similarity between languages is evaluated based on analysis of a small set of pre-defined key words. Each language is represented by a set of 10 key words. Our hypothesis is that it is possible to perform meaningful clustering of European languages using phonetic similarity between just 10 common words from each language. This clustering is performed via SOM.

Selecting a suitable word set is not an easy task, because human languages constantly evolve, and the meaning of words may be culture-

dependent. For example, the meaning of the word 'cool' commonly used in modern English is quite different from its meaning in 16th century English. Many words, e.g., 'computers' and 'information', have been introduced very recently. Such words cannot be used for comparative analysis, because they have not been used in older versions of contemporary languages. So, we need to select words that have culture-independent meaning that is preserved over thousands of years. It may be reasonable to expect that similarity between such 'stable' words reflects similarity between languages. In this study, the key word set includes digits from 'one' to 'ten' in different European languages. Phonetic representation of these words for digits goes back to pre-historic times, so their phonetic similarity can be used as a criterion for similarity between languages. However, measuring phonetic similarity (between words in different languages) requires expert linguistic knowledge needed for phonetic transcription of words. So in this study, the phonetic similarity (between words in different languages) will be measured indirectly by their alphabetic similarity. In order to compare words from different languages, they need to have common representation. Hence, we choose only European languages based on the Latin alphabet (or its minor variations). These 18 European languages are: English, Norwegian, French, Flemish, Czech, Portuguese, Slovakian, Finnish, Estonian, Swedish, Danish, Dutch, German, Spanish, Italian, Polish, Hungarian, and Croatian. For these languages, we used English transcription of words (numbers) obtained from a public-domain web site http://www.zompist.com/euro.htm. Table 6.1 shows the spelling of the numbers 'one' to 'ten' for eighteen European languages.

Each word is represented by the first 3 letters in this word, regardless of the order of appearance of these letters. Each word is encoded as a 27-dimensional feature vector with binary 0/1 values, corresponding to 26 letters of English alphabet and a *blank_space* symbol. Each coordinate of this vector corresponds to a particular letter in English alphabet. The presence of a letter in a word is indicated by the value 1 in the corresponding position. For example, the word 'one' is encoded as a 27-dimensional vector where entries # 15, 14 and 5, corresponding to these three letters, are set to 1, and all other entries are zeros. See Fig. 6.22. The same encoding is used for all 10 word-numbers in each language, and these ten 27-dimensional binary vectors are concatenated, to produce a 270-dimensional feature vector encoding for each language. The standard batch SOM algorithm is then applied to eighteen 270-dimensional feature vectors encoding European languages. Clustering of these 18 samples was performed using standard Euclidean distance in the

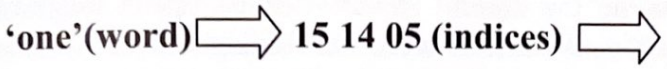
270-dimensional input space. The batch SOM method used the following user-defined parameters:

- two dimensional map with 4x4=16 units;
- initial neighborhood size = 1, final neighborhood size = 0.15;
- total number of iterations = 70.

The final trained SOM is shown in Fig. 6.23. It shows that European languages mapped onto each SOM unit form clusters consistent with known linguistic classification. For instance, this map clearly shows clusters (units) containing Slavic, Germanic and Romance languages. It also identifies the well-known linguistic similarity between Finnish and Estonian languages.

TABLE 6.1 Spelling of numbers in 18 European languages.

English	one	two	three	four	five	six	seven	eight	nine	ten
Norwegian	en	to	tre	fire	fem	seks	sju	atte	ni	ti
Polish	jeden	dwa	trzy	cztery	piec	szesc	sediem	osiem	dziewiec	dziesiec
Czech	jeden	dva	tri	ctyri	pet	sest	sedm	osm	devet	deset
Slovakian	jeden	dva	tri	styri	pat	sest	sedem	osem	devat	desat
Flemish	ien	twie	drie	viere	vuvve	zesse	zevne	achte	negne	tiene
Croatian	jedan	dva	tri	cetiri	pet	sest	sedam	osam	devet	deset
Portuguese	um	dois	tres	quarto	cinco	seis	sete	oito	nove	dez
French	un	deux	trois	quatre	cinq	six	sept	huit	neuf	dix
Spanish	uno	dos	tres	cuatro	cinco	seis	siete	ocho	nueve	dies
Italian	uno	due	tre	quattro	cinque	sei	sette	otto	nove	dieci
Swedish	en	tva	tre	fyra	fem	sex	sju	atta	nio	tio
Danish	en	to	tre	fire	fem	seks	syv	otte	ni	ti
Finnish	yksi	kaksi	kolme	nelja	viisi	kuusi	seitseman	kahdeksan	yhdeksan	kymmen en
Estonian	uks	kaks	kolme	neli	viis	kuus	seitse	kaheksa	uheksa	kumme
Dutch	een	twee	drie	vier	vijf	zes	zeven	acht	negen	tien
German	erins	zwei	drie	vier	funf	sechs	sieben	acht	neun	zehn
Hungarian	egy	ketto	harom	negy	ot	hat	het	nyolc	kilenc	tiz



0	0
0	1
0	2
0	3
0	4
1	5
0	6
0	7
0	8
0	9
0	10
0	11
0	12
0	13
1	14
1	15
0	16
0	17
0	18
0	19
0	20
0	21
0	22
0	23
0	24
0	25
0	26

FIGURE 6.22 Encoding first 3 letters of a word in a 27-dimensional vector.

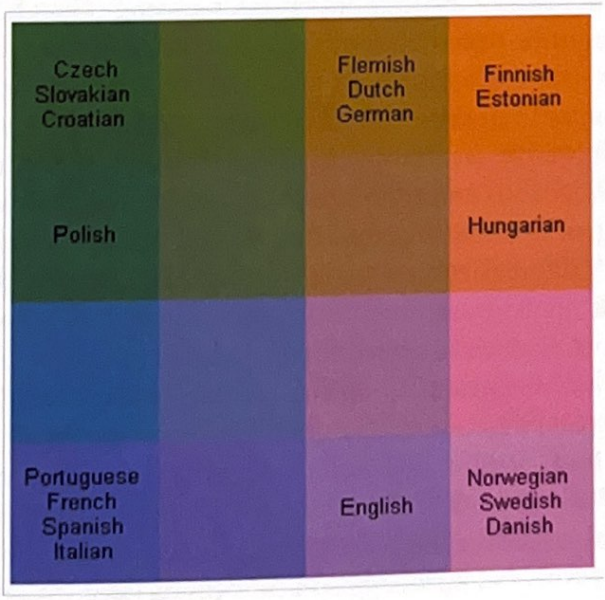


FIGURE 6.23 Clustering of European languages obtained using 4x4 SOM. The Neighboring SOM units are represented by similar color.

Good clustering of very complex objects (natural languages) achieved via simple analysis of a small word set is rather surprising. In fact, most linguists will be highly suspicious of such a simplistic approach and dismiss its clustering results as a pure coincidence. These concerns are addressed next. The simplicity and power of this approach are based on the proper selection of a stable word set. The possibility of correct clustering using the digits data set is evident from the visual inspection of data shown in Table 6.1. SOM method simply provides a formal mechanism for revealing the natural clusters (of European languages). However, the existence of such a word set hiding correct similarity relationships between European languages is indeed surprising. This methodology can be further validated using a different stable word set, i.e., ten words denoting body parts (head, ear, eye, nose, etc.). See Problem 6.18. Applying the same SOM clustering approach to this data set yields correct clustering of European languages, similar to results shown in Fig. 6.23.

6.8 SUMMARY AND BIBLIOGRAPHIC NOTES

Presentation of neural network methods in this chapter follows the predictive learning framework. That is, neural network methods are regarded as nonlinear estimators implementing the ERM or SRM approach using the squared loss function. This view enables better understanding of such methods and various heuristics for model complexity control. Historically, neural network algorithms have been introduced informally, using biological motivation and terminology. See (Haykin 1999) for a comprehensive description of neural networks. Understanding of statistical properties of neural network methods came later, in mid-1990's (Cherkassky and Mulier 1997, Ripley 1996).

Initial work on artificial neural networks dates back to the early days of computing. A mathematical model of a single neuron (in Section 6.1) was proposed by McCulloch and Pitts (1943). In early 1960's, this model was used for classification (Rosenblatt 1962) and for regression tasks (Widrow and Hoff 1960). The delta rule, described in Sectrion 6.1, was proposed by Widrow and Hoff (1960) for training a single linear neuron (called Adaline). Extension of the delta rule to training multilayer networks, known as backpropagation, appeared later (Werbos 1974) and it led to a renaissance in ANN research in late 1980's and early 1990's. Note that different nonlinear optimization techniques (other than backpropagation) can be used for minimizing the squared error