ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Applications of l_p-Norms and their Smooth Approximations for Gradient Based Learning Vector Quantization

M. Lange[1], D. Zühlke[2], O. Holz[3], and T. Villmann[1]

1- University of Appl. Sciences Mittweida - Dept. of Mathematics
Mittweida, Saxonia - Germany

2- Fraunhofer-IAIS - Dep. of Organized Knowledge
Sankt Augustin, Germany

3- Fraunhofer-ITEM - Hannover , Germany

Abstract. Learning vector quantization applying non-standard metrics became quite popular for classification performance improvement compared to standard approaches using the Euclidean distance. Kernel metrics and quadratic forms belong to the most promising approaches. In this paper we consider Minkowski distances (l_p-norms). In particular, l_1-norms are known to be robust against noise in data, such that, if this structural knowledge is available in advance about the data, this norm should be utilized. However, application in gradient based learning algorithms based on distance evaluations need to calculate the respective derivatives. Because l_p-distance formulas contain the absolute approximations thereof are required. We consider in this paper several approaches for smooth consistent approximations for numerical evaluations and demonstrate the applicability for exemplary real world applications.

1 Introduction

Utilization of non-standard (non-Euclidean) metrics is one of the key ideas in learning vector quantization to improve the performance of classification learning. Whereas in traditional vector quantization the Euclidean distance is standard, kernel methods like support vector machines make use of kernel similarities [20]. Other examples are weighted Euclidean distances [8], general bilinear forms [22], correlations [23], functional norms and Sobolev distances [14, 18], divergences [4, 24] or kernel distances [20, 25], to name just a few.
Recently, l_p-norms with $p \neq 2$ became popular as alternative dissimilarities in machine learning approaches [1, 6, 16]. Depending on the parameter p, l_p-norms show different behavior, which makes them interesting for many applications [2, 17]. For example, the larger the p-value, the greater is the influence of noise for the receptive p-norm. Thus, for noisy data l_1-norms are more appropriate than the Euclidean norm [7].
Yet, differentiation of general l_p-norms and their induced dissimilarity measures suffers from the inconsistency for the origin $\mathbf{x} = \mathbf{0}$ due to the inherent absolute value calculation. Therefore, the application of l_p-norms in gradient based machine learning approaches requires smooth consistent approximations of the derivatives.
The aim of the paper is twofold: First we consider several smooth approximations of dissimilarities induced by l_p-norms and semi-norms and provide consistent approximations of their derivatives. Thus, they can immediately be applied in gradient based learning vector quantization (GLVQ,[19]). Second, we show in two real world examples the successful application of l_1-norms for LVQ.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

2 l_p-Norms, derivatives and approximations thereof

2.1 Basic notations and properties

In this section we provide useful approximations of the l_p-norms such that derivatives become available also at the origin $\mathbf{x} = \mathbf{0}$. For this purpose, we consider the Minkowski l_p-norm $\|\mathbf{x}\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}$ for $1 \leq p < \infty$ with the corresponding *Minkowski distance* $d_p^* (\mathbf{v}, \mathbf{w}) = \|\mathbf{v} - \mathbf{w}\|_p$. The frequently used quantity

$$d_p (\mathbf{v}, \mathbf{w}) = \left(\|\mathbf{v} - \mathbf{w}\|_p \right)^p \tag{1}$$

is only a dissimilarity measure violating the triangle-inequality and the linearity of distances. However, it is frequently considered in machine learning.
The choice of the p-value causes different behavior of the dissimilarities. The larger p the more important great variations become in a single dimension. For $p < 1$ small variations, are emphasized and the unit 'circle' becomes concave, see Fig.1.

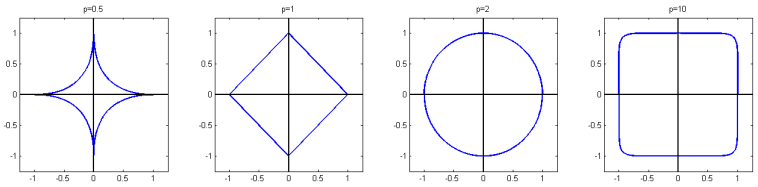

Figure 1: Unit circles for several Minkowski-p-norms $\|\mathbf{x}\|_p$: from left to right $p = 0.5$, $p = 1$ (Manhatten), $p = 2$ (Euclidean), $p = 10$.

For $0 < p < 1$, d_p^* is only a *quasi-norm* [15] fulfilling *p-triangle inequality* $\|\mathbf{v}\|_p^p + \|\mathbf{w}\|_p^p \leq \|\mathbf{v} + \mathbf{w}\|_p^p$. However, now (1) is a *translation-invariant* distance [11].
The *weighted dissimilarities*

$$d_p^{\lambda} (\mathbf{v}, \mathbf{w}) = \sum_{i=1}^{n} \lambda_i |v_i - w_i|^p \text{ and } d_p^{\Omega} (\mathbf{v}, \mathbf{w}) = \sum_{i=1}^{m} \left(\left| [\Omega (\mathbf{v} - \mathbf{w})]_i \right| \right)^p \tag{2}$$

with $\lambda_i \geq 0$ subject to $\sum_{i=1}^{n} \lambda_i = 1$ and $\Omega \in \mathbb{R}^{m \times n}$, respectively, were proposed for $p = 2$ in relevance learning for learning vector quantization [8, 22]. Thereby, we denoted $[\mathbf{x}]_i = x_i$.

2.2 Formal derivatives for $d_p (\mathbf{v}, \mathbf{w})$

Utilization of those dissimilarities with in gradient based machine learning methods like self-organizing maps or GLVQ requires the calculation of the derivatives $\frac{\partial d_p (\mathbf{v}, \mathbf{w})}{\partial \mathbf{w}}$, where $\mathbf{v} \in \mathbb{R}^n$ is a data vector $\mathbf{w} \in \mathbb{R}^n$ is a prototype. Supposing $0 < p < \infty$, the formal derivative is

$$\frac{\partial d_p (\mathbf{v}, \mathbf{w})}{\partial w_k} = -p \cdot |z_k|^{p-1} \cdot \frac{\partial |z_k|}{\partial w_k} \tag{3}$$

with $\mathbf{z} = \mathbf{v} - \mathbf{w}$. The gradient (3) can be written in vectorized form as

$$\frac{\partial d_p (\mathbf{v}, \mathbf{w})}{\partial \mathbf{w}} = -p \cdot |\mathbf{z}|^{\star (p-1)} \circ \frac{\partial |\mathbf{z}|}{\partial \mathbf{z}} \tag{4}$$

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

where $\mathbf{x} \circ \mathbf{y} = (x_1 \cdot y_1, \ldots, x_n \cdot y_n)^\top$ denotes the Hadamard product. Further, $\mathbf{x}^{\star k}$ denotes the componentwise power of $\mathbf{x} = \left(x_1^k, \ldots, x_n^k\right)^\top$ and $\frac{\partial |\mathbf{z}|}{\partial \mathbf{w}} = -\frac{\partial |\mathbf{z}|}{\partial \mathbf{z}}$ holds. Analogously, we find

$$\frac{\partial d_p^\lambda(\mathbf{v},\mathbf{w})}{\partial \mathbf{w}} = -p \cdot \lambda \circ |\mathbf{z}|^{\star(p-1)} \circ \frac{\partial |\mathbf{z}|}{\partial \mathbf{z}} \text{ and } \frac{\partial d_p^\Omega(\mathbf{v},\mathbf{w})}{\partial \mathbf{w}} = -p \cdot \Omega^\top \left(|\mathbf{s}|^{\star(p-1)} \circ \frac{\partial (|\mathbf{s}|)}{\partial \mathbf{s}} \right) \tag{5}$$

with $\mathbf{s} = \Omega \mathbf{z} \in \mathbb{R}^m$. We observe that we need the derivative of the absolute value function, which has to be approximated for the origin.
For $p = \infty$ the above dissimilarity formulas (2) involve the maximum function $\max(\mathbf{x}) = \max_i(x_i)$. Thus we get

$$\frac{\partial d_\infty^\lambda(\mathbf{v},\mathbf{w})}{\partial \mathbf{w}} = -\frac{\partial \max(\lambda \circ |\mathbf{z}|)}{\partial \mathbf{z}} \text{ and } \frac{\partial d_\infty^{\mathbf{\Omega}}(\mathbf{v},\mathbf{w})}{\partial \mathbf{w}} = -\Omega^\top \frac{\partial \max(|\mathbf{s}|)}{\partial \mathbf{s}} \tag{6}$$

analogously. Here, the derivative of the maximum function is required. The derivatives for the relevance weights λ_i and the $\Omega_{k,j}$ in case of the weighted dissimilarities (2) are obtained analogously for relevance or matrix learning and can be found in [13].

2.3 Approximations for the functions $\max(\mathbf{x})$ and $|\mathbf{x}|$

In the following we consider smooth approximation for the function $\max(\mathbf{x})$ and $|\mathbf{x}|$ as well as their derivatives.
At least two variants for the the maximum function easily are well-known: The first is the α-*softmax* function

$$\mathcal{S}_\alpha(\mathbf{x}) = \frac{\sum_{i=1}^n x_i e^{\alpha x_i}}{\sum_{i=1}^n e^{\alpha x_i}} \tag{7}$$

with $\alpha > 0$, frequently applied in optimization and neural computation [3, 9]. A value $\alpha < 0$ in (7) yields a smooth minimum approximation whereas for $\alpha \to 0$ a soft approximation of the mean is obtained. The second frequently applied variant is the α-*quasimax* function

$$\mathcal{Q}_\alpha(\mathbf{x}) = \frac{1}{\alpha} \log\left(\sum_{i=1}^n e^{\alpha x_i} \right) \tag{8}$$

proposed by J.D. COOK [5]. This α-*quasimax* can be seen as a kind of a *generalized functional mean* or *quasi-arithmetic mean* discussed in [12]. One can easily verify that $\mathcal{Q}_\alpha(\mathbf{x}) \leq \max(\mathbf{x}) + \frac{\log(n)}{\alpha}$ is always valid .
A smooth approximation of the absolute value function was proposed in [21]:

$$|x|_\alpha = (x)_\alpha^+ + (-x)_\alpha^+ \tag{9}$$

with $(y)_\alpha^+ = \max(\mathbf{y}_0)$ for $\mathbf{y}_0 = (y, 0)^\top$. CHENG&MANGASARIAN proposed a convex α-approximation $(y)_\alpha^+ = y + \frac{1}{\alpha}\log(1 + e^{-\alpha y})$. One verifies that this is equivalent to

$$(y)_\alpha^+ = y + \mathcal{Q}_\alpha(\mathbf{y}_0) \tag{10}$$

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

using the α-*quasimax* function $\mathcal{Q}_\alpha$ from (8) [13]. Inserting these formulas in (9), we obtain an approximation

$$|x|_\alpha^{\mathcal{Q}} = \frac{1}{\alpha} \log\left(2 + e^{-\alpha x} + e^{\alpha x}\right) \tag{11}$$

referred as α-*quasi-absolute*. Hence, $|x|_\alpha^{\mathcal{Q}}$ is consistent with $\mathcal{Q}_\alpha(\mathbf{x})$ with the upper bound $||x| - |x|_\alpha^{\mathcal{Q}}| \leq 2\frac{\log(2)}{\alpha}$ [17].
Alternatively, we can replace in (10) the α-*quasimax* function $\mathcal{Q}_\alpha$ by the α-*softmax* function $\mathcal{S}_\alpha$ which leads to

$$|x|_\alpha^{\mathcal{S}} = \frac{x \cdot \left(e^{\alpha x} - e^{-\alpha x}\right)}{2 + e^{\alpha x} + e^{-\alpha x}} \tag{12}$$

denoted as α-*soft-absolute*.

2.3.1 Approximation of the derivatives

If the above introduced smooth approximations are used in gradient based numerical methods, neural networks or other methods in machine learning the derivatives have to be known. We provide the respective formulas in the following: In particular, we obtain

$$\frac{\partial |x|_\alpha^{\mathcal{Q}}}{\partial x} = \tanh\left(\frac{\alpha}{2}x\right) \text{ and } \frac{\partial |x|_\alpha^{\mathcal{S}}}{\partial x} = \tanh\left(\frac{\alpha}{2}x\right) + \frac{\alpha x}{2\left(\cosh\left(\frac{\alpha}{2}x\right)\right)^2}$$

for the α-*quasi-absolute* $|x|_\alpha^{\mathcal{Q}}$ and for the α-*soft-absolute* $|x|_\alpha^{\mathcal{S}}$, respectively. Although $|x|_\alpha^{\mathcal{Q}}$ and $|x|_\alpha^{\mathcal{S}}$ look quite different, their derivatives differ only slightly by the additive deviation term

$$\Delta_{\mathcal{SQ}}(\alpha x) = \frac{\alpha x}{2\left(\cosh\left(\frac{\alpha x}{2}\right)\right)^2} \,. \tag{13}$$

For the α-softmax function $\mathcal{S}_\alpha(\mathbf{x})$ from (7), the gradient can be expressed in terms of $\mathcal{S}_\alpha(\mathbf{x})$ itself

$$\frac{\partial \mathcal{S}_\alpha(\mathbf{x})}{\partial x_k} = \frac{e^{\alpha x_k}}{\sum_{i=1}^n e^{\alpha x_i}} \left[1 + \alpha\left(x_k - \mathcal{S}_\alpha(\mathbf{x})\right)\right] \tag{14}$$

whereas for the α-quasimax function $\mathcal{Q}_\alpha(\mathbf{x})$ from (8) we simply obtain

$$\frac{\partial \mathcal{Q}_\alpha(\mathbf{x})}{\partial x_k} = \frac{e^{\alpha x_k}}{\sum_{i=1}^n e^{(\alpha x_i)}} \,. \tag{15}$$

Again, we observe only a slight variation $\frac{\partial \mathcal{S}_\alpha(\mathbf{x})}{\partial x_k} = \frac{\partial \mathcal{Q}_\alpha(\mathbf{x})}{\partial x_k} \cdot \nabla_{\mathcal{SQ}}$ with the multiplicative corrector $\nabla_{\mathcal{SQ}} = [1 + \alpha\left(x_k - \mathcal{S}_\alpha(\mathbf{x})\right)]$.

3 Applications

We applied the dissimilarity $d_p^\Omega(\mathbf{v}, \mathbf{w})$ from (2) in the matrix variant of GLVQ (GMLVQ,[22]) for two different datasets. The first one is a set of 4120 tiling microarray (data dimension $n = 24$) of corresponding to exonic and intronic/intragenic regions in chromosome 3 of *C.elegans*. Tiling micro array data usually contain a lot of noise

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

such that learning of these data is difficult. A detailed description of the data can be found in [1]. In the investigation described in this publication, the standard LVQ1 algorithm was applied for prototype learning, which is based on the (weighted) Euclidean norm $d_2^\lambda(\mathbf{v}, \mathbf{w})$. However, inherent prototype selection as well as relevance parameters λ_i optimization were done using the l_1-distance $d_1^\lambda(\mathbf{v}, \mathbf{w})$ to deal with the noisy data. Thus, there is an inconsistency in the approach. The best performance was obtained for 6 prototypes per class: 89.3%. In our application we used the same number of prototypes (6 per class). We applied GMLVQ with $d_p^\Omega(\mathbf{v}, \mathbf{w})$ for $p = 1$ and $p = 2$ consistently used for prototype learning and matrix adaptation in a 4-fold cross-validation. The achieved test accuracies are 90.8% and 88.8%, respectively. Thus consistent l_1-learning improves the result. Further, we did not recognized any significant difference regarding the used approximation (α-*quasi-absolute* or α-*soft-absolute*, $\alpha = 20$).
In a second application we analyzed spectral data obtained from a gas chromatography–mass spectrometry (GC-MS) analysis of volatile organic components in exhaled breath for detection of inflammatory processes in the lung. The GC-MS spectra were delivered as 334-dimensional spectra covering a measurement time interval of $20\mathrm{min}$. The dataset contains 48 spectra partitioned into two classes. A detailed data description can be found in [10]. We used only one prototype per class in GMLVQ and conducted a 8-fold cross-validation. Again we applied $d_p^\Omega(\mathbf{v}, \mathbf{w})$ for $p = 1$ and $p = 2$ consistently for prototype learning and matrix adaptation. We achieved the test accuracies 85.4% and 81.3%, respectively. The better l_1-result clearly emphasizes this choice of dissimilarity. This is in agreement with the knowledge about the noise influence for GC-MS spectra regarding the peak height [26] and the above mentioned robust behavior of l_p-distances for smaller p-values also reported in [7].

4 Conclusion

In this paper we consider l_p-distances and discuss smooth approximation thereof, which are required when used in gradient based learning methods based on dissimilarity evaluations between data and reference vectors. Generally, l_p-distances contain the absolute value function, which causes difficulties for the (numerical) calculation of the derivatives. We provide smooth approximations and explain the respective derivatives. We compare the utilization of the weighted l_1-distance with the weighted Euclidean distance for two applications. The first one is also in comparison to an earlier but inconsistent approach. We demonstrate that l_1-distances can be successfully applied using the described approximation techniques.

Acknowledgment

The authors gratefully acknowledge very helpful discussions with MARC STRICKERT, UNIVERSITY MARBURG and MICHAEL BIEHL, UNIVERSITY GRONINGEN.

References

[1] M. Biehl, R. Breitling, and Y. Li. Analysis of tilling microarray data by learning vector quantization and relevance learning. In H. Yin, P. Tino, E. Corchado, W. Byrne, and X. Yao, eds., *Proc. Intelligent Data Engineering and Automated Learning (IDEAL)*, LNCS 4881, pages 880–889. Springer, 2007.

[2] M. Biehl, M. Kästner, M. Lange, and T. Villmann. Non-Euclidean principal component analysis and Oja's learning rule – theoretical aspects. In P. Estevez, J. Principe, and P. Zegers, editors, *Advances in Self-Organizing Maps: 9th International Workshop WSOM 2012 Santiage de Chile*, volume 198 of *Advances in Intelligent Systems and Computing*, pages 23–34, Berlin, 2013. Springer.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press,, 2004.

[4] A. Cichocki and S.-I. Amari. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12:1532–1568, 2010.

[5] J. Cook. Basic properties of the soft maximum. Working Paper Series 70, UT MD Anderson Cancer Center Department of Biostatistics, 2011. http://biostats.bepress.com/mdandersonbiostat/paper70.

[6] O. Golubitski and S. Watt. Distance-based classification of handwritten symbols. *International Journal on Document Analysis and Recognition (IJDAR)*, 13(2):133–146, 2010.

[7] Z. Gu, M. Shao, L. Li, and Y. Fu. Discriminative metric: Schatten norms vs. vector norm. In *Proc. of The 21st International Conference on Pattern Recognition (ICPR 2012)*, pages 1213–1216, 2012.

[8] B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15(8-9):1059–1068, 2002.

[9] S. Haykin. *Neural Networks. A Comprehensive Foundation*. Macmillan, New York, 1994.

[10] O. Holz, A. Gaida, S. Schuchardt, B. Lavae-Mokhtari, H. biller, M. Rosano, J. Hanrahan, and J. Hohlfeld. Volatile organic compounds (VOC) in exhaled breath after experimental ozone exposure. In *Proc. of the American Thoracic Society 2013 International Conference*, page A4101. American Thoracic Society, 2013.

[11] I. Kantorowitsch and G. Akilow. *Funktionalanalysis in normierten Räumen*. Akademie-Verlag, Berlin, 2nd, revised edition, 1978.

[12] A. Kolmogorov and S. Fomin. *Reelle Funktionen und Funktionalanalysis*. VEB Deutscher Verlag der Wissenschaften, Berlin, 1975.

[13] M. Lange and T. Villmann. Derivatives of l_p-norms and their approximations. *Machine Learning Reports*, 7(MLR-04-2013):43–59, 2013. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_04_2013.pdf.

[14] J. Lee and M. Verleysen. Generalization of the l_p norm for time series and its application to self-organizing maps. In M. Cottrell, editor, *Proc. of Workshop on Self-Organizing Maps (WSOM) 2005*, pages 733–740, Paris, Sorbonne, 2005.

[15] E. Pekalska and R. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*. World Scientific, 2006.

[16] B. Póczos, S. Kirshner, and C. Szepesvári. REGO: Rank based estimation of Rényi information using Euclidean graph optimization. In *Proc. of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9 of Journal of Machine Learning Research (JMLR), 2010.

[17] M. Riedel, F. Rossi, M. Kästner, and T. Villmann. Regularization in relevance learning vector quantization using l_1-norms. In M. Verleysen, editor, *Proc. of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2013)*, pages 17–22, Louvain-La-Neuve, Belgium, 2013. i6doc.com.

[18] F. Rossi, N. Delannay, B. Conan-Gueza, and M. Verleysen. Representation of functional data in neural networks. *Neurocomputing*, 64:183–210, 2005.

[19] A. Sato, K. Yamada, and J. Tsukumo. A multi-template learning method based on LVQ. In *Proc. ICNN'93, International Conference on Neural Networks*, volume II, pages 632–637, Piscataway, NJ, 1993. IEEE, IEEE Service Center.

[20] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

[21] M. Schmidt, G. Fung, and R. Rosales. Fast optimization methods for l1 regularization: A comparative study and two new approaches. In J. Kok, J. Koronacki, R. Mantaras, S. Matwin, D. Mladenič, and A. Skowron, editors, *Machine Learning: ECML 2007*, volume 4701 of *Lecture Notes in Computer Science*, chapter 28, pages 286–297. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[22] P. Schneider, B. Hammer, and M. Biehl. Adaptive relevance matrices in learning vector quantization. *Neural Computation*, 21:3532–3561, 2009.

[23] M. Strickert, F.-M. Schleif, U. Seiffert, and T. Villmann. Derivatives of Pearson correlation for gradient-based analysis of biomedical data. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, (37):37–44, 2008.

[24] T. Villmann and S. Haase. Divergence based vector quantization. *Neural Computation*, 23(5):1343–1392, 2011.

[25] T. Villmann, S. Haase, and M. Kästner. Gradient based learning in vector quantization using differentiable kernels. In P. Estevez, J. Principe, and P. Zegers, editors, *Advances in Self-Organizing Maps: 9th International Workshop WSOM 2012 Santiage de Chile*, volume 198 of *Advances in Intelligent Systems and Computing*, pages 193–204, Berlin, 2013. Springer.

[26] T. Villmann, F.-M. Schleif, M. Kostrzewa, A. Walch, and B. Hammer. Classification of mass-spectrometric data in clinical proteomics using learning vector quantization methods. *Briefings in Bioinformatics*, 9(2):129–143, 2008.